

Population stability within reach?

There is a chance that next week's UN conference on population and development at Cairo will give a necessary fillip to declining fertility around the world, but only if the red herrings on the agenda can be avoided.

THE good news about the world's population is that fertility is declining everywhere. The average number of children born to women during their reproductive lifespan, called the total fertility rate, is estimated to have fallen by more than 10 per cent between the second half of the 1970s and the second half of the succeeding decade. In most of the rich countries of the world, the average number of children born during a reproductive lifetime is already below the magic number of 2.1, estimated to be that required to replace the existing population. In Italy, the Vatican's hinterland, births are now outnumbered by deaths and fertility is only 60 per cent of replacement level, but even in the poor parts of the world, fertility fell by more than 13 per cent between the late 1970s and late 1980s. That is, indeed, good news. It shows that what is called the 'demographic transition' (from high birth and death rates to low rates of both) is under way worldwide.

The bad news is that the shift is too slow for comfort. In much of Western Europe in the nineteenth century, it took 60 years or more for the demographic transition to be completed. In Britain, for example, birth and death rates comparable with those in developing countries now were falling perceptibly by 1811, and had attained modern levels by the early 1870s; much of the population growth since then has been the result of greater longevity. The causes of these huge changes are well-understood: rapid industrialization, urbanization, growing prosperity, beginnings of universal education and the reduction of infant mortality following the improvement of public health. That is the background to the Third International Conference on Population and Development, organized by the United Nations (UN) at Cairo for 5–13 September. But can the world afford to wait for 60 years or more for that transition to come about in the natural course of events?

Stability

The answer to the question is "no!". That, at least, is the theme running through the draft of the final statement of the conference now being circulated as a vehicle for the noisy debates that will occupy the next two weeks. Although fertility is falling, the absolute number of people alive is rising quickly. Eventual stability is not unattainable, but the question is that of how many more than 7 billion people there would then be. And while there is now no cause to fear that shortages of food and other material necessities of life would be threats to the mere survival of even such large numbers, there is no doubt that a world so crowded would remain

plagued by the uncivil inequalities between rich and poor that now mar too many people's lives. It would be less well managed than it might be.

Containment

The UN's draft would have the forthcoming conference agree on all the right things. It recognizes that the containment of population growth is both a precondition for successful economic development and a consequence thereof. Population stability is therefore part and parcel of the process of development. The agenda is rich in the kinds of contradictions with which the aid community is familiar. Poor peasant families have too many children because they know from experience that many will die in infancy, but are then so impoverished that they cannot afford the clean water supplies that would prevent infant death from intestinal diseases, for example. How to break these interlocking vicious circles? The UN draft gives pride of place to three goals: equality with men for women everywhere, the improvement of public health to reduce infant and maternal mortality and basic education for all. The technology of contraception, while still important, is a secondary issue. Population containment is a social science problem.

That liberal declaration is unexceptionable, which does not mean that it will be generally endorsed at Cairo. How many governments will give more than a formal nod in the direction of the most ringing declaration of women's rights that the UN has yet produced? There are also more contentious issues for Cairo, not least the unavoidable connection between population policies and the human right to freedom. The draft says that the coercion of individuals should play no part in a government's population policy, but what will the Chinese make of that? (In China, the 'one family, one child' policy has led to a total fertility rate of 1.6 and to a male:female sex ratio estimated at 1.6:1, presumably by selective abortion and infanticide.) And where is the line to be drawn between legitimate public education and coercive propaganda?

The issue of abortion will also be contentious, but in reality is a red herring. The Vatican has advertised its wish to see the final declaration at Cairo insist that abortion should not be an instrument of population policy. Its views deserve respect. All governments should regard abortion as the least satisfactory method of birth control, and should even be prepared to say so. But in the most populous countries, China and India, abortion has been legitimized for



the purpose, and nothing said in Cairo will change that. It would be out of tune with the principles of the UN Charter, which does not allow intervention in the domestic affairs of member states, to make a fuss on this issue.

That, sadly, is not the only red herring on the agenda. In the best UN tradition, the drafting committees have set out to cure all the world's ills at Cairo. The result is that the draft declaration is laden with contentious issues not strictly relevant to population stability. Sustainable development is a recurring theme. There are even specific recommendations on the way in which the debts of developing-country governments should be handled. It is to be hoped that the minor rows that these will generate will not so distract the Cairo conference that the opportunity for strengthening the impending global demographic transition will be lost. □

Courteous Linus Pauling

One pity of Linus Pauling's death is that even vitamin C could not keep him living longer.

THE death last week of Linus Pauling at 93 is a sad business for everybody. He was one of the now few surviving links with that heroic age when modern science was fashioned from the bare essentials of quantum mechanics as made plain in the 1920s at Copenhagen, Göttingen and Heidelberg. With a few deft monographs, and in particular, with *The Nature of the Chemical Bond*, he induced a whole generation of young people to take seriously the notion that the properties of molecules could be calculated. By carrying the message of quantum mechanics to the New World, he did for chemistry what Oppenheimer did for physics.

But quantum chemistry, for Pauling, was merely incidental to his chief interest, the exercise of his unparalleled sense of three-dimensional geometry in the understanding of how molecules, especially large molecules, are constructed. Watson and Crick were understandably anxious, in the early 1950s, that Pauling, from an earlier generation, might be the first to get the structure of DNA. It is still a mystery that he did not. Versatility was his strong suit.

And iconoclastic independence. There is great relish in the tale of Pauling's arrival at a tuxedo dinner at President John F. Kennedy's White House straight from the lines of an anti-nuclear demonstration outside the railings. Yet he was not an embittered iconoclast, but one of great charm and courtesy. And who, having seen him lecture on some topic close to his scientific interest, will not still be infected with his enthusiasm for understanding, and can have forgotten those great arms waving like semaphores to tell the world what fun it all is?

Pauling's last exercise in iconoclasm, his advocacy of megadoses of vitamin C as a defence against cancer and other ills, is in a different case. The charitable judgement of that episode is that Pauling never really buckled down to accepting the complexities of the evidence usually required to sustain a hypothesis in biology. (Even so, there is no reason why the BBC should have labelled him a heretic —

see below.) But it is important that concern with the effects of vitamin C did not blind Pauling to the rest of science. When quasi-crystals were discovered, he was one of the first with an explanation. To the end, last week, he remained alert, seeing the fun in things. He will be missed. □

Heresy at the BBC

The British Broadcasting Corporation deserves no credit for a meretricious series of television films.

THE British Broadcasting Corporation (BBC), not to be mistaken for an agency of the British government, is one of the world's outstanding broadcasting organizations. Over the years, it has helped to carry authentic news of world events to parts of the world deprived thereof, and has entertained and instructed its domestic audience in a manner that is often admired (and imitated). On the whole, it has done well by science. BBC television, for example, invented the idea that working scientists could be made to explain what they were about to intelligent interlocutors in longish programmes consisting of several weekly parts.

Now, and sadly, the BBC is scrambling like many other British institutions to 'reposition' itself in British society and in the world. (Although the government has recommended that the BBC's right to collect a 'licence fee' from every British owner of a television receiver, which is where the money comes from, parliament has still to debate the issue.) Presumably as part of that exercise, the BBC has just broadcast a remarkable series of television programmes under the rubric of "Heretic" which does science and the BBC itself no service.

The idea is simple, and intrinsically entertaining. Invite an iconoclast to outline an eccentric theory, and then invite more orthodox people to say why they do not agree. The iconoclast and his critics never meet face to face. The result is that both the iconoclast and his (or her, but there were no women heretics in the BBC's sample of six) critics are made to seem foolish, the former because of the inevitable obsessiveness that shows through, the latter because disembodied statements of the general opinion inevitably sound pompous. Good fun is had by all.

The BBC's iconoclasts included the authors of 'cold fusion', Dr Jacques Benveniste (who believes that water has a memory, and with whom *Nature* has had several brushes), and the late Linus Pauling (see above). Dr Rupert Sheldrake (who holds that morphogenetic fields condition the world we live in) was another. The programmes asked, in effect, why science will not take seriously the questions asked by manifestly honest people. The answer is, of course, that the iconoclasts have always failed to demonstrate that their questions are meaningful. But the concept of the 'good question' would no doubt have diminished the general audience for these programmes. Yet when the renewal of the BBC's charter is eventually debated, this meretricious exercise will no doubt be quoted as another public service done. □

Western inferno provokes a lot of finger-pointing, but little action

Washington. Ecologists and foresters say that 28 huge and ferocious forest fires now raging in the western United States are part of a worsening trend to be blamed on a century of vain attempts to suppress forest fires in the region. Tangible proposals to reverse those mistakes have been drawn up, but deep divisions between environmentalists, foresters and local people make them unlikely to be implemented.

Five million acres of forest are expected to be lost to fire before the summer is out — the second largest loss this century — despite a \$1-billion federal effort employing 150 helicopters, 40 air tankers and 20,000 fire-fighters to douse the flames. Many people see the fires as a 'natural' consequence of drought. But ecologists and foresters agree that these intense fires are not 'natural' at all, but rather the result of a century of land mismanagement that has suppressed regular forest burning and transformed millions of acres of forest into what one recent report called "an explosive time bomb waiting to ignite".

As research reveals more about the natural fires and fires set by native Americans which maintained the forests in the past, a consensus has developed between environmentalists and foresters that the forests urgently need thinning. Most even agree that logging and prescribed fire should both play a role. But bad blood between the two groups means that the fight for money to pay for the new strategy has been in vain — even as the

every sixteen years between 1716 and 1893 — it is not known if Indians or lightning did the burning — but never since.

When European gold-diggers first arrived a hundred years ago, the Boise forest contained about 100 trees an acre — 70 per cent fire-resistant ponderosa pine and 30 per cent Douglas fir, says Mealey. Today a typical acre has more than 500 trees, 70 per cent of them the fire-prone Douglas fir. Forests throughout the western United States are now "at high risk of catastrophic wild-fire", he says.

Defending his Forest Service predecessors from environmentalists' charges of mismanagement, Mealey says "the science wasn't on the table" until the past decade of dry weather exposed the risk. "We were in a moist period, so the symptoms weren't apparent," he explains. Dry weather from 1986 exposed the problem: "after that the ecological wheels started to come off".

Science has the tools to get the wheels back on, but the task is seriously complicated by fighting between the mechanics. "There's a general consensus on why the forests are more susceptible to burning today," says Jim Agee, a forest ecologist at the University of Washington, Seattle. "But when we look to future solutions, the consensus breaks down because it involves a lot more than biology."

Everyone agrees that the forests need thinning. Regular, prescribed burning could keep them thinned, but logging of smaller trees is needed first, to stop the fires burning out of control. The environ-

mental lobby does not trust the foresters to log only small trees. And local people — ranging from animal lovers to real-estate developers — mistrust prescribed burning. The result is an expensive and dangerous impasse.

Mark Rey, vice-president of the American Forest and Paper Association, says his organization supports both techniques, seeing little difference between them in practice. The real problem is political, he believes. "Environmental activists are unable

to separate the issue from their antipathy towards logging and forest management," he says. "They'd prefer the issue not be raised, because at its heart is the fact that you can't conserve the forest by walking away from it."

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Boise National Forest — a tinderbox of 500 trees an acre.

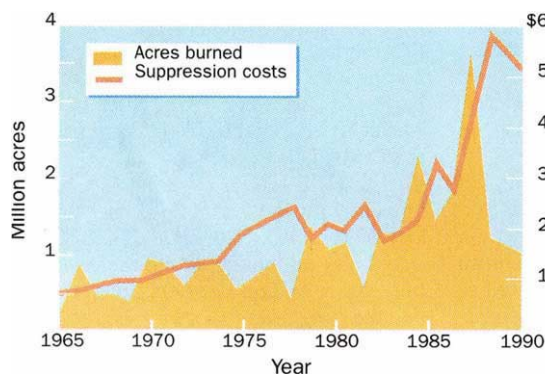
Not surprisingly Jon Roush, the president of the Wilderness Society, rejects this accusation. "I don't know of any mainstream environmental group that would oppose good, strong forest management," he says. The main resistance to change comes from within the federal agencies such as the Forest Service, he contends, adding that "there are people in the forestry industry who see the issue as an opportunity to push for more logging — and to put environmentalists in a bad light".

The warning of a 'time bomb' was issued by a National Commission on Wildfire Disasters, in a strongly worded report delivered earlier this year. The report contains a thorough action plan and a call for further research on the economics of the forest management options it proposes.

The commission was chaired by Neil Sampson, vice-president of the conservation group American Forests. Sampson is trying to build support in Washington for the approach outlined in the report. "Unfortunately, when the flames die down we'll be hard pressed to keep attention on this," he says.

But that may not matter. From the point of view of Congress, the \$1 billion likely to be spent this year on fire-fighting is 'non-discretionary' spending, and as such counts for little. Funds for a much smaller programme to prevent fire would need to be appropriated from valuable discretionary funds. As one unnamed congressional staff member told Sampson's bemused commission, "we pay agencies to fight fires, not light them".

Colin Macilwain



cost of the huge annual fire-fighting operation soars.

Steve Mealey is an old-time forester who supervises the 2.5-million-acre Boise National Forest in Idaho, where 2,000 fire-fighters are trying to control this year's blazes. He has seen the average annual loss by fire at Boise grow from 3,000 acres a year before 1985 to more than 50,000 acres since.

A cut-away of an aged ponderosa pine tree in Mealey's office explains why. The tree has rings showing that it was scorched

Committee looks into the future of US space science

Washington. A newly formed panel of the US National Research Council (NRC) begins work this month to recommend a new course for space science in an era of shrinking budgets, increased cooperation and greater demand that research be tied to economic goals.

The Future of Space Science panel, chaired by John Armstrong, former head of research at IBM, will also consider the possibility of creating a new national institute of space science within the National Aeronautics and Space Administration (NASA), modelled loosely on the National Institutes of Health (NIH).

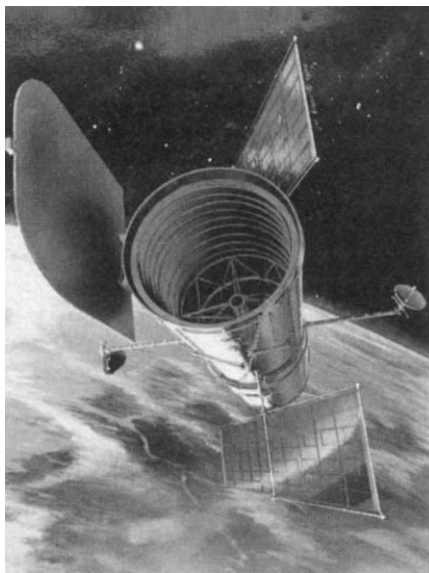
The Senate subcommittee that decides NASA's budget ordered the study last September in the wake of the agency's decision to divide its office of space science into three smaller offices responsible for Earth sciences, life and microgravity sciences, and astronomy and planetary exploration.

The subcommittee felt that this new structure would make it difficult for NASA to establish cross-disciplinary priorities, and for Congress to monitor its activities. It was also "deeply troubled" by the lack of plans (meaning money) for new space science missions after 1997, something the new NRC panel will also be looking at.

According to NASA's chief scientist, France Cordova, the agency already has a plan to address this problem. If its science managers can cut \$100 million from their overall operational budget, they will receive a funding 'wedge' of up to US\$300 million for new science missions. The money would be trimmed from other NASA programmes. The incentive is working, according to Cordova, who says that managers of the Cassini Saturn mission have already identified "pretty dramatic proposed savings" in their operations budget.

The problems go beyond just money, however. In fact, NASA's space science budget is bigger than the entire annual budget of the National Science Foundation. But the money is not always well spent, according to another recently published NRC report. Entitled *A Space Physics Paradox*, the report has the intriguing subtitle, "Why has increased funding been accompanied by decreased effectiveness in the conduct of space physics research?"

The report's authors found that "factors such as planning, marketing, the funding process, and project management have become as responsible for the increased delays, costs, and frustration levels in space physics as technical complications relating to increasing project size and complexity". They recommend that the balance of funding be shifted toward smaller programmes,



Will there be space science after Hubble?

even if it means reducing the number of large programmes. And although they looked at only one discipline — space physics — the report's authors believe "the subject and results of this study apply to many other areas of science as well".

Another difficulty facing space science is that its role in the US space programme is not well defined. "We have not established an overall space policy for the country," says Louis Lanzerotti of AT&T's Bell Laboratories, who until recently chaired the Space Studies Board. The last attempt to define such a policy — the 1990 report by the presidentially commissioned Augustine Committee — recommended that science should be NASA's top priority. That advice was largely ignored as the White House and the space agency concentrated instead on getting the space station approved.

NASA is often accused of over-emphasizing human flight to the detriment of science. But although shuttle and space station managers do sometimes treat science as an afterthought — the NRC's Space Studies Board wrote to Daniel Goldin, administrator of NASA, last month to complain that Spacelab science experiments were being bumped to make room for logistics payloads — space science and the human space programme are more often symbiotic. The shuttle and space station need onboard research to justify their existence. And without those projects to draw political, industrial and public support, funding for the rest of NASA's space science activities would almost certainly dwindle. The new NRC panel will have to take that reality into account as it tries to chart the future of space science.

Tony Reichhardt

New president for South Korea's university of steel

Tokyo. One of South Korea's newest universities, the Pohang University of Science and Technology (POSTECH), last week appointed a US-trained Korean engineer as its new president following the sudden death of POSTECH's first president, Hogil Kim, in a sporting accident earlier this year.

POSTECH is unusual in that it receives a substantial part of its funds from one of the world's most profitable steel manufacturers, the Pohang Iron and Steel Company (POSCO) (see *Nature* 364, 380; 1993). Now ranked as one of South Korea's leading universities, POSTECH was established with POSCO's support on a plush new campus in 1987. The company provides about 60 per cent of the university's annual research funds of about US\$28 million via an adjacent company research institute, the Research Institute of Science and Technology.

POSTECH has recruited nearly all its faculty from top universities and research organizations in the United States. The new president, Sooyoung Chang, is no exception. Chang received his PhD from the University of Maryland in 1971 and taught at the State University of New York before working as a systems engineer in private companies in the United States.

David Swinbanks

Japan's H-II rocket fizzles on launch pad

Tokyo. Officials of Japan's National Space Development Agency (NASDA), who hope to compete in the world market for the commercial launch of satellites, were embarrassed last week when their new H-II rocket failed to get off the launch pad in its first mission.

The H-II, which was successfully test-launched for the first time earlier this year (see *Nature* 367, 500; 1994), fired its main liquid fuel engine six seconds before lift off, but then solid fuel boosters failed to ignite and the launch was aborted. The rocket was intended to place a two-tonne test communications satellite into geostationary orbit.

NASDA has traced the problem to a faulty computer chip in ground-based launch control equipment. The H-II is much more expensive than competing commercial launch vehicles like Europe's Ariane rockets and NASDA officials have been emphasizing the reliability of their technology, rather than price. But last Thursday's failure came on top of an earlier launch cancellation due to a faulty valve. A third attempt to launch the rocket will be made on Sunday (28 August).

David Swinbanks

Tax laws impede joint project by US and Japan

Tokyo. Japan's tax laws are standing in the way of a pioneering attempt by Japan and the United States to give their computer designers access to the most advanced optoelectronic devices under development in company and government research laboratories in both countries.

As part of Japan's Real World Computing programme, a 10-year US\$700-million initiative by the Ministry of International Trade and Industry (MITI) to develop new massively parallel computers, the United States and Japan signed an agreement in June to establish the US-Japan Optoelectronics Project (JOP), which is expected to begin in October.

The project provides a broker service in each country to bring together suppliers of new optoelectronic devices under development in the laboratory on the one hand, and designers of advanced computer systems who will evaluate the devices in experimental prototype machines on the other. (These designers could work in private or government laboratories, or in universities.)

In Japan, the broker system is run by the Optoelectronics Industry and Technology Development Association. In the United States a broker system will be set up shortly by the National Institute of Standards and Technology.

Japan is a recognized world leader in optoelectronics, and government officials on both sides expect the flow of devices to be predominantly from Japan to the United States. But tax regulations in Japan are making it very difficult for Japanese companies to provide US users with devices.

Companies in Japan are given a tax break on their research and development expenditure to encourage them to spend even more. Twenty per cent of any annual increase in a company's research and development expenditure is tax-deductible — but only if the company research institutes do not engage in any commercial activity.

Under the broker system, users pay for the devices to which they are given access and also enter into agreements on intellectual property rights. But the Japanese companies producing the devices in the laboratory cannot sell them without losing their research and development tax privileges.

To avoid infringing tax laws, US and Japanese government officials have agreed that, although the devices must be the latest technology coming out of company laboratories, they can be provided by the laboratory to a commercial arm of the company, then sold by this unit to the JOP broker.

But this has led to other problems, says Kazuo Kyuma of Mitsubishi Electric's central research laboratory in Osaka. Kyuma has been approached by a delegation of US government officials from various agencies, including the Department of Defense,

seeking access to a novel artificial retina chip, developed by Kyuma, that detects and processes images in much the same way as the human eye.

Mitsubishi cannot yet mass-produce the retina chip in its factories because the device is still under development and not ready for large-scale production. Kyuma can produce the devices in his laboratory and transfer them to a Mitsubishi factory for sale to US users through the JOP broker system, but the company is concerned about the quality of the laboratory-produced devices.

Kyuma says the situation in Japan is quite different from that in the United States where company laboratories often produce and sell devices. In Japan, company research institutes never do this because of tax regulations and they are thus not geared up to produce reliable devices for sale.

Kyuma believes that the most likely so-



lution is that his laboratory will supply the devices to US users through a Mitsubishi factory with the understanding that they may not have the same reliability as factory-made components.

David Swinbanks

India set to end 'gene robbery'

New Delhi. International drug companies will no longer have free access to India's genetic wealth. A tough new law, which comes into force next year, will prevent foreign companies from helping themselves to commercially valuable genes.

The new legislation names more than 9,000 medicinal plants, in addition to industrially important microbes, that will no longer be allowed to leave India. It will be illegal to export crop germ plasm without a bilateral 'exchange' agreement with the state-owned national gene bank.

Officials from the Ministry for Environment and Forests (MEF) say that in the absence of any regulatory mechanisms, foreign agencies, particularly drug companies, have been plundering India's genetic resources for their own benefit.

They cite *Rauwolfia serpentina* as an example. This plant provides raw material for an antihypertensive drug whose annual sales are US\$260 million in the United States alone, but India does not get any share of the profit, they say. And, the American Type Culture Collection in Maryland has more than 380 fungal and 90 bacterial accessions from India, some of them patented.

Speaking recently at a meeting of the Indian Academy of Sciences on protection of intellectual property rights, Palpu Pushpangadan, director of the tropical botanical gardens in Trivandrum, claimed that US and Japanese companies had often taken away soil samples, along with soil microorganisms, from different ecosystems in India, without any form of control.

Pushpangadan, who is also president of the National Society of Ethnopharmacology, has data on 7,500 medicinal plants that have

been used as folk remedies. But to prevent foreign companies from exploiting these plants, his research institute in Trivandrum — under instruction from MEF — is to withhold the information from electronic databases to which foreigners have access.

The international scramble for India's genetic diversity came into focus last week when two Germans, claiming to be tourists, were caught trying to smuggle more than 30,000 insects of various types out of Delhi airport. They are awaiting charges.

MEF officials say the new legislation will end what they call "gene robbery" and ensure that India gets compensation from companies that market products using India's genetic material.

Controlling the country's genetic wealth has become crucial, particularly after the conclusion last year of the General Agreement on Tariffs and Trade, which allows patenting of genes and microorganisms. In addition, India is a signatory to the biodiversity convention which gives states sovereign rights over their biological resources. This is why India has the right to restrict access to its gene pools, say officials.

India has also taken steps recently to get back some of the materials it has already lost. Two months ago, the International Rice Research Institute in Manila in the Philippines agreed to return to India 5,000 accessions of traditional rice strains from Assam state, which were taken away by the institute in the 1960s. The United States will similarly be asked to return duplicates of all crop germ plasm it has collected from India between 1950 and 1970.

K. S. Jayaraman

UK to set up DNA database of criminals . . .

Paris. Britain is to set up the world's first national database of DNA profiles from convicted criminals, putting DNA alongside fingerprints among the forensic scientist's frontline tools. The United States is expected to follow suit by the end of the year.

The decision reflects an emerging consensus that the controversies over the reliability of DNA profiling are little different from those of other forensic tests, and that they will be quickly resolved, not through public debate, but by defence lawyers.

Most observers also agree that a database of limited DNA profiles of convicted offenders poses no new ethical problems. They point out that it would be little different from databases of fingerprints or mugshots, and may provide a major new 'deterrent' against violent crimes.

The UK civil rights group Liberty says it opposes a national database for offenders, because it would "be open to abuse", but it accepts a database limited to DNA from sex offenders and murderers. It also wants guarantees that individuals would have the right

to check their stored profiles, and that courts would not convict individuals on DNA evidence alone.

Under the British scheme — announced last week by Michael Howard, the Home Secretary — police could take DNA samples from anyone charged with an offence punishable by imprisonment. Their profiles would be digitalized on optical discs, and be routinely screened against samples of DNA taken from crime scenes. The database would cost around £28 million in the first year.

One area of uncertainty is whether the police will keep DNA profiles from acquitted suspects on file. According to Howard, removing such profiles may be technically difficult because they would be stored on the same disc as other samples from a case.

Senior police officials are also concerned that DNA samples will not be taken from those previously convicted of serious crimes, unless they offend again following their release. This means the database will be ineffective until it has been in operation for

several years. But introducing retrospective legislation would pose legal problems, said Howard.

The forensic science service is expected to provide Howard with full details of how the database will operate within the next few weeks. The final plans will be opened to consultation for three months, before becoming law under the Criminal Justice and Public Order Bill, which is expected to pass through parliament before the end of the year.

The US Federal Bureau of Investigation (FBI) is also about to launch a national database of DNA profiles. Eight states have already begun, and between them hold 24,000 DNA records of convicted offenders, according to an FBI spokesperson. Several suspects have already been identified and convicted of serious crimes using the state databases.

Some scientists, however, remain concerned over the reliability of forensic DNA profiling techniques in general, and the use of databases in particular. Daniel Hartl, a population geneticist at Harvard▶

. . . while US may regulate DNA testing laboratories

San Francisco. As analysts examine DNA evidence in the O. J. Simpson murder case, US legislators are considering regulating the laboratories performing such tests. A crime bill now being debated in Congress would place forensic DNA testing under the regulatory authority of the Federal Bureau of Investigation (FBI) and set standards for licensing of laboratories. It would require proficiency testing twice a year by an accredited company and would set aside \$250,000 for a feasibility study of national blind proficiency testing.

Some forensic DNA experts believe that laboratory error is underestimated in courtroom debates over the technology (see above). Defence lawyers for Simpson, who has been charged with murdering his ex-wife and another man, have already begun homing in on the performance record of Cellmark Diagnostics, which is analysing DNA evidence in the case.

Cellmark has been accused of sloppy analysis because of two false matches the Maryland company made during a 100-sample proficiency study administered by the California Association of Crime Laboratory Directors (CACL D) in the late 1980s. Critics point out additional errors made in a report to the CACL D that were later corrected. Cellmark admits that the first report was confusing but says it was not in error.

Industry-wide results from that study and a later analysis by Collaborative Testing Service suggest an overall false-positive

rate of one per cent to four per cent of reported matches, estimates Jonathan Koehler, assistant professor at the University of Texas at Austin.

But Cellmark argues that its error rate is a mere 0.5 per cent on the basis of the original proficiency test plus another 300 samples analysed in the intervening years. The company now conducts its own tests as well as subscribing to several testing programmes, said Mark Stolorow, director of operations for Cellmark. "In the last five years there have been no documented errors either in our proficiency tests or in our case work," he said.

Defence lawyers, however, criticize Cellmark's tests for lack of scientific rigour. In general, errors result from misinterpretation of data, ambiguous data and mixing of samples, says William Thompson, associate professor at the University of California at Irvine.

The crime bill would ask the FBI to develop standards for performance and for proficiency testing. This would result in greater acceptance of the technology in the courtroom, says Stolorow.

The majority of courts in the United States do accept DNA evidence, but a few states, including Arizona, Massachusetts and Minnesota, continue to express concern. In California, where the Simpson case is being tried, appellate courts have been backing away from their original enthusiasm.

In November 1992, the California Su-

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

O. J. Simpson: his DNA may decide trial.

preme Court upheld an appellate panel's ruling in a San Francisco case, *People versus Barney/People versus Howard*, that statistical methods for evaluating DNA evidence remained in scientific dispute. Therefore, the court said, such statistical procedures were inadmissible, as was the DNA evidence itself.

While appellate courts have followed the Barney decision, lower courts generally continue to admit DNA typing. A newer method based on PCR technology is being tested for the first time at the appeals court level in a sexual assault case scheduled for a decision very soon.

The high-profile Simpson case is certain to influence acceptance of DNA analysis by the public and in the courtroom, both prosecution and defence lawyers say. This comes at a time when the National Research Council is preparing to re-examine the use of DNA typing (see *Nature* 367, 101; 1994).

Sally Lehman

►University, says courts need to guard against the "likely" risk that a search of tens of thousands of profiles would yield a random match between an entry and a crime sample.

Hartl says that courts need to confirm such matches by checking a fresh sample from the suspect against a larger number of genetic loci. This would also check for errors in data entry, he adds. Hartl agrees, however, that the profiling procedures used by courts have generally improved. And "DNA-savvy" defence lawyers have made courts more aware of potential sources of errors, he says.

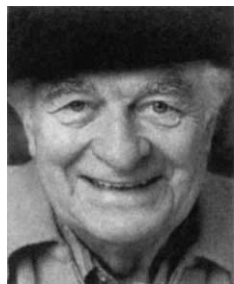
The controversy over the probabilities of a match occurring by chance began in 1991, when Hartl and Richard Lewontin — now also at Harvard — argued that the FBI's use of only three reference US populations — Caucasians, Blacks and Hispanics — did not take into account the presence of numerous ethnic subgroups.

Simply put, this would mean that if an Amish committed a murder, for example, then DNA from another innocent Amish suspect would more closely resemble that found at the scene of the crime than would DNA from the Caucasian, Black or Hispanic reference groups. In more usual circumstances, the same logic would apply to Polish, Italian and other ethnic subgroups in the United States.

Lewontin and Hartl's arguments gave defence lawyers scientific ammunition, and sent the FBI scurrying around the planet to collect data on the genetic make-up of ethnic subgroups (see *Nature* 355, 663; 1992). According to Hartl, however, the FBI has not yet changed its practices in the light of these data. "They argued that the differences in frequencies between ethnic groups didn't matter. Their data show that they do, and it's very inconvenient."

Hartl argues that the entire problem of statistics could be avoided by using eight probes instead of the three or four now commonly used. He says that eight — as now used in the state of Minnesota — would "make it inconceivable, by any scientific standards, that the blood came from someone else". But Cellmark, a major DNA fingerprinting company based in the UK, defends the four locus 'standards' as "sufficiently discriminatory". **Declan Butler**

Linus Pauling dies



Nobel prize winner and anti-nuclear campaigner Linus Pauling died last week at the age of 93. See leader article, page 584.

Royal Society of Canada spurned as national academy

Ottawa. A panel appointed in 1992 by Canada's former Progressive Conservative government has recommended against making the 100-year-old Royal Society of Canada the country's national academy.

According to its report, submitted to the present Liberal government last April but released to the public only on 11 August, "the Society does not have the organizational capacity to evolve into the kind of Academy envisioned by the panel". The report says that in the panel's view such an academy should "represent the perspectives of all sectors in the economy, generate funds to ensure its independence from government, and achieve and maintain the public profile required to stimulate and contribute to public debate of complex issues".

Instead of using the Royal Society, the panel recommended setting up a new body, independent from government but funded initially by up to C\$250,000 a year, and led by a "limited number of highly regarded Canadians". Its "pivotal role" would be "to identify, analyse and provide advice on issues of critical importance to Canadians" — issues as diverse as "the impact of technology on jobs, ethics and biotechnology, and sustainable development".

The mandate of the panel, chaired by Brian Segal, publisher of the popular news magazine, *Maclean's*, and including two Royal Society fellows, was to advise the federal government on whether it should support the formation of a national academy and whether the Royal Society could serve in that purpose.

The society, founded in 1883 and granted its royal charter by Queen Victoria, was modelled on the Royal Society of London and l'Institut de France. As an independent, non-profit-making organization, it promotes learning and research in the arts and sciences. Its fellows (numbering around 1,400) are elected on the basis of distinction in their field. Its usual operating budget has been around C\$500,000 annually, from subscriptions and other sources.

The society has for some time considered itself as Canada's national academy: its descriptive brochure, published in 1990, is subtitled *Profile of a National Academy*, and the brochure describes its role as such.

In 1989 the government's department of industry provided a grant of C\$5 million over five years to enable the society to strengthen its organization, management and fund-raising capacity, apparently with the aim of further developing its "academy" role. Under the terms of the grants, it was also to produce three studies that would evaluate Canadian research, consider the advancement of women in scholarship and

promote public awareness of science.

But on 24 December last year, the department terminated the final C\$250,000 instalment of the grant without warning. This led to the dismissal of all but three of the society's core staff of about a dozen.

An assessment of the society's use of the grant, led by a consultant appointed by the federal department, was included in the panel's report. It said that less of the money than intended was used to fund the studies, mainly because the society failed to obtain the expected revenues from other sources. Moreover, it was not happy about the studies themselves. It characterized the society's work in research evaluation as "poor", and in the advancement of women as "medium" in quality and usefulness. The third study is continuing.

The society has reacted defensively to the panel's report. It already performs many of the proposed new academy's functions, it says. It has, for example, mounted expert panels for studies of issues such as the impact of AIDS, at the request of the government, which have been well received.

It accuses the government of "trying to reinvent the wheel", and warns that the creation of a new academy could cause confusion, particularly abroad because "many national academies abroad recognize the Royal Society of Canada as their Canadian counterpart... and are perfectly happy with this arrangement".

The proposal that the head of the academy would be recruited and paid by the industry department would be regarded with suspicion, it argues. Additionally, it could run into difficulties with Quebec's French-speaking intellectuals, many of whom are nationalists: "a new synthetically created pan-Canadian organization, fostered by the federal government, could well arouse suspicion (and cause the abstention) of many Quebec scholars, artists and scientists", while the society has dealt with the two cultures for more than a century.

The society also points out that during the five years during which it received the industry department grant, it raised more than C\$6 million from other sources, which was used to fund activities appropriate to a national academy.

While the panel claims that its new academy would "build on the strengths of existing organizations" and "maximize the impact of scarce resources", observers note ironically that some of its functions were formerly carried out by the Science Council of Canada, which was abolished by the same Conservative government that created the panel.

David Spurgeon

Brazilian science searches for new direction

São Paulo. Three separate studies of the state of science and technology in Brazil have provoked soul searching among the country's scientists. The studies all call for increased financing of applied research, and for increased links with private enterprise.

But doubts still linger about the wisdom of dumping a model based on state-sponsored basic research that has achieved good results in the recent past, even though it is recognized that this model may no longer be appropriate in a world characterized by lower trade barriers and increased pressure to achieve international competitiveness.

The studies were commissioned by the Ministry of Science and Technology at the request of international agencies such as the World Bank. Published in May, near the end

of the current government's term (elections are in October), they set a framework for the science policies of its successor.

"We need to stop the decline in resources, stop the brain drain, make efforts to work with private enterprise, and stop big projects that have not worked, such as the nuclear energy deal with Germany," says Simon Schwartzman, a researcher at University of São Paulo and Getlio Vargas Foundation, who coordinated one of the studies.

There are already signs that the government acknowledges the dangers of cutting back support for research. In one of its last acts before the election, it has increased funding for the Ministry of Science and Technology by 6 per cent next year, at a time when almost all other ministries have exper-

iened cuts of between 16 and 20 per cent.

The largest of the three studies, costing US\$1.6 million and spawning a 6,000-page document, focused on the competitiveness of Brazilian Industry, and was coordinated by economist Luciano Coutinho, of the State University of Campinas (Unicamp) in São Paulo State.

The second study, coordinated by Schwartzman, dealt with the current status of several areas of science and technology in Brazil, notably in universities and research institutes. The third, dealing with the institutional framework for financing research, was compiled by Francisco Biato of the Ministry of Economy.

Speaking at a meeting organized by the Getlio Vargas Foundation — a ►

Universities in bid to take over research institute

London. A consortium of three universities this week unveiled plans to take over the Natural Resources Institute (NRI), a major government research institute, just a few months after the government's multi-departmental efficiency scrutiny recommended that transfer of research institutes to universities should be considered "routinely" (see *Nature* 370, 89; 1994).

Under the proposal, put to the Overseas Development Administration (ODA) earlier this month, the NRI's active research arm and the laboratories at Chatham, southern England, would be taken over by the University of Greenwich. The marketing and business side of the NRI would become a separate non-profit company jointly owned by the three colleges making up the consortium, the University of Greenwich, the University of Edinburgh and Wye College, part of the University of London.

The announcement comes as the University of Greenwich is in the process of moving its School of Earth Sciences to NRI's site in Chatham. The NRI had been seeking a sub-tenant since 1992 to rent excess office and laboratory space. The university and NRI will offer joint courses from September.

Other research institutes are watching this move, and any subsequent developments on the ownership of the NRI, with interest, in light of the recommendations by the scrutiny team, which are in a four-month consultation period until November.

Earlier this month the ODA, the current owners of NRI, announced Phase

II of a special ownership study to investigate further the options for ownership of NRI. Phase I of the study, commissioned by the ODA in January, included the possibility of the transfer of NRI to a university.

The NRI, which specializes in research in natural resource management in developing countries, including resource assessment and pest management, was formed in 1987 by the amalgamation of the Tropical Development and Research Institute and the Land Resources Development Centre. It was made an agency of the ODA in 1990 as part of the government's plans to prepare some public-sector research institutions for privatization.

Rob Black, who is chairman of the ODA branch of the Institution of Professionals, Managers and Specialists (IPMS), the main union of the 530 NRI staff, said that union members had yet to be consulted on the plans put forward by the consortium, which were only made public this week. But he did expect that NRI staff would be asking the ODA for assurances that job losses would be kept to a minimum and that there would be

some continuity in terms of the naming of the institute, which has undergone several changes in its recent history.

Staff are right to be sceptical, according to Valerie Ellis, assistant general secretary of the IPMS. Although the union is in favour of university links, there may be hidden dangers because universities are not as close to research institutes as may be imagined, she says.

"The major problem is that NRI's mission is about helping developing countries, including technology transfer, and that's why it's a government laboratory and not a university department already," argues Ellis.

Despite claims by Sir Peter Levene, head of the efficiency scrutiny team, that rationalization does not necessarily mean job cuts, staff are concerned that phase II of the ownership study will mean more redundancies on top of the 47 already this year.

"I think everybody accepts that NRI will diminish in size, [but] I still think most people probably would prefer to remain in the civil service as an agency, that they do their job best in that way," says Black.

Lady Chalker, minister for overseas development, will ultimately make the decision on ownership, probably by the end of this year. But NRI is quite sure of its preferred option. "I think there's no doubt at all that it's going to be quite difficult to find an ownership option which has as many of the attractions of merging with a university," says Anthony Beattie, chief executive of NRI.

Maggie Verrall



The University of Greenwich will share NRI's site at Chatham in Kent.

►business and public administration school — in São Paulo, Michael Gibbons, director of the Science Policy Research Unit at the UK University of Sussex, summarized the situation by pointing out that “the scientific, technological and educational developments established in the 1970s no longer seem to be working”.

The three papers indicate that the main reason is economic. Most of the major institutions associated with the development of a robust national scientific and technological capability — in particular government research establishments and universities — are in financial difficulty, and their links with industry remain weak.

There is a general consensus that the private sector should do more to support science. Only 15 per cent of the country's bill for research and development is paid by the private sector, compared to around 70 per cent in some Asian countries such as Taiwan, Korea and Japan.

Yet how to attain the competitiveness of the Asian ‘tigers’ remains hotly debated. Many at the Getlio Vargas meeting felt that this could still be achieved through State backing for basic science.

But others pointed out that India, for example, is a Third World country that has invested heavily in science with only limited results in terms of economic development; they suggested instead that Brazil copy the market-driven, low-to-high-tech approach of countries such as Taiwan and South Korea.

Whichever path is taken, one hurdle remains: the antiquated Brazilian state. Any science and technology policy requires a stable economy, if it is to be successful, and in the case of Brazil this may need long-term reform of the state.

Brazil still lacks a unified science policy. Ivan Rocha, of the University of Brasilia, who participated in Biato's study, says that science funding agencies need to link science to broader industrial and educational goals; otherwise funding agencies will continue “shooting in all directions”, he says.

The next government will have to deal with the big projects of the past, including schemes for a nuclear submarine, military aircraft and a space rocket launcher. Military research still accounts for one-fifth of the government's spending on research.

Some are demanding reductions in such spending. But others, such as Geraldo Lesbat Cavagnari Filho, a former Army colonel, who is now a researcher at Unicamp, says the government should finish what has already been started.

Cavagnari has given this advice to the presidential candidate now second in public opinion polls, Luiz Inacio Lula da Silva, of the left-wing Workers Party (Partido dos Trabalhadores). Many scientists may not like it. But, as Cavagnari says, they have much to lose, as 90 per cent of the armed forces' researchers are civilians.

Ricardo Bonalume Neto

Koop may set up new centre for alternative medicine

Washington. Joseph Jacobs, who steps down in October as head of the Office of Alternative Medicine at the US National Institutes of Health (NIH), may be about to join forces with C. Everett Koop, the outspoken surgeon general in the Reagan administration, in setting up a centre for alternative medicine at Dartmouth Medical College in New Hampshire.

The Koop Institute at the college promotes curriculum changes in medical schools. David Serra, a co-director of the institute, confirmed rumours that Koop was having discussions with Jacobs but refused to say about what. He did say that if such a centre went ahead its aim would be to investigate alternative therapies with the assumption that some regimens work. But, he added, “we can't get too far ahead of ourselves because there is some opposition within Dartmouth”.

Jacobs and Koop both believe that primary care could play a greater role in health care in the United States (see *Nature* 370, 501; 1994). Jacobs, who is reticent about his future plans, says that it is in this context that he would like to see alternative medicine given more serious attention.

Jacobs has had a rough ride as the first head of the Office of Alternative Medicine, which was established by a congressional mandate to further research into practices such as acupuncture and herbalism. Colleagues say that he has been caught between the traditional biomedical community and extremists in alternative medicine.

“My critics at both ends of the spectrum are caught in a mind set of 20 years ago,” says Jacobs. “A way forward for alternative medicine is through primary health care. The critics and zealots don't really understand that. They are in a political battle and are using the NIH as a battlefield,” he added.

Jacobs believes that the Office of Alternative Medicine should not be attached to the National Institutes of Health, which he says is a “highly specialized, quintessential research organization” that is not suited to the more mundane research needed in alternative medicine. Nevertheless, he believes that to dismiss conventional medicine is narrow-minded. Like many at the centre of the debate, Jacobs would be more comfortable if alternative medicine were called complementary medicine (as it is in much of Europe).

Steve Martin, assistant professor of medicine and of the history and philosophy of medicine at the Albert Einstein

College of Medicine in New York, says, “the term alternative medicine encompasses a wide array of techniques, some compatible with conventional medicine, some not”.

If Koop and Jacobs do establish a centre for alternative and complementary medicine at Dartmouth, they will join a small

group of respected medical schools, such as Harvard and Columbia, that have set up similar centres within the past year or so.

In fact, a paper published in the *New England Journal of Medicine* last year by David Eisenberg, who established

Harvard's centre, has provided some of the impetus behind the increasing interest of medical students and faculty in alternative practices. Eisenberg concluded from a telephone survey that, in 1990, Americans spent about US\$13.7 billion on alternative therapies.

One of the most widely used of these alternative therapies was acupuncture. The Office of Alternative Medicine has collected data from clinical trials that lawyers representing acupuncturists were to present this week as part of a petition to have acupuncture needles removed from the Food and Drug Administration's (FDA)'s list of investigational devices.

Investigational devices should be used only experimentally, but, in practice, acupuncture needles are widely used for treatment. James Turner, the lawyer presenting the petition, expects a decision from the FDA within the next six months. If acupuncture needles are taken from the list of investigational needles, it will be possible for patients to reclaim the cost of treatment from federal programmes that pay for the health care of the elderly and the poor.

The next item on the agenda of the Office of Alternative Medicine is to persuade the FDA to allow some clinical trials of Chinese herbs. One application was submitted by Freddie Kronenberg, a physiologist who heads Columbia's centre for alternative and complementary medicine. She wants to investigate the effectiveness of Chinese herbs for the relief of hot flushes during the menopause.

Helen Gavaghan

IMAGE
UNAVAILABLE
FOR
COPYRIGHT
REASONS

Acupuncture gains credibility as an alternative therapy.

Jeremy Hartley/Panos Pictures

Science and human rights

SIR — We should like to amplify your News story (*Nature* 368, 680; 1994) about a session of the Committee on International Freedom of Scientists (CIFS) at the annual meeting of the American Physical Society. Your report detailed one case but other important issues were also discussed at the session.

For example, the Chinese physicists Wang Juntao and Liu Gang were among those imprisoned after peacefully participating in the demonstration in Tiananmen Square in 1989 and have been held under severe conditions. Wang's wife, Hou Xiaotian, spoke at the meeting of his critical medical condition. CIFS had joined other groups in petitioning for Wang's early release to receive medical treatment for hepatitis B. Six days after the CIFS session (and just before President Bill Clinton's decision to renew China's 'Most Favored Nation' status), Wang was 'medically paroled' and is now being treated in the United States. Liu Gang remains in Linguan Prison No. 2 under harsh conditions.

As a result of protests by CIFS and other scientific and human rights organizations, charges were dropped against Vil Mirzayanov, the Russian whistleblower who revealed continued secret Russian manufacture of binary chemical weapons in spite of a treaty prohibiting such manufacture. In Uzbekistan, mathematicians and physicist members of Birlik, a pro-democracy group, were arrested and imprisoned for peaceful expressions of opinion. Abdumammob Pulatov provided the meeting with details of the latest cases of repressed scientists.

CIFS has also protested at the US government's infringement of the right of scientists to travel freely to and from Cuba to exchange scientific information. A statute dating back to the First World War — the Trading with the Enemy Act — effectively prevents Americans from attending conferences in Cuba, even those sponsored by the International Union of Pure and Applied Physics and the International Council of Scientific Unions. Similarly, Cuban scientists are not normally admitted to the United States; as a result, an international conference was shifted to Vienna last year.

On the complaint of a former East German physicist that your correspondent referred to, the physicist concerned admitted being an *inoftizial Mitarbeiter* (unofficial collaborator) of the Stasi (the East German secret police), but claimed he had not received fair treatment at the hearing that resulted in his dismissal. Over the past two years, CIFS has been in touch with his colleagues and academic superiors and officials of the German Physical Society to determine whether he received

a fair hearing. The case is still in progress.

Another case concerns Mohammed Al-Mass'ari, a Saudi Arabian professor of physics at Riyadh University, who was arrested, imprisoned and held incommunicado while his office and home effects were confiscated. Many of his colleagues were also arrested. He is a member of the Committee for the Defence of Legitimate Rights.

In these and other cases, CIFS interventions are intended to express concern for the right of individual physicists and other scientists to pursue their scientific activities in accordance with internationally recognized principles.

Joseph Birman

Fang Lizhi

Herman Winick

Eugene Chudnovsky

Yuri Orlov

Committee on the International

Freedom of Scientists 1994,

American Physical Society,

One Physics Ellipse,

College Park, Maryland 20740-3844, USA

AIDS research

SIR — Bernard N. Fields makes a number of cogent comments in his proposal for redirection of AIDS research (*Nature* 369, 95–96; 1994). But to suggest that "it is no longer reasonable to invest major resources in the search for drugs and vaccines" is a highly provocative point of view, one whose implementation might affect the lives of hundreds of thousands of HIV-infected people over the next 10 years.

This discourse is not solely an academic debate about what gets studied: it is at least equally about who gets federal grant money. The two points are inseparable, and the latter cannot help but affect beliefs about research priorities. Many basic laboratory scientists not directly associated with AIDS work have complained bitterly about the perceived shift in funding to AIDS and other disease-specific research. It is essential that widespread public review should precede any changes in priorities as proposed by Fields. Most critically, we must hear from those scientists who have spent much of the past decade directly working on this problem. In the absence of a clear, specific plan and an open debate, it will be all too easy to see any change as an effort to shift grant support from one group of scientists to another.

Letters submitted for Correspondence should be typed, double-spaced, on one side of the paper only.

The best way to accomplish productive change would be to convene a formal (and long overdue) assessment process of AIDS research, one which includes but is independent from the US National Institutes of Health as well as the interests of any individual scientific specialty. Fields should indeed be heard, but so should the views of many other scientists who have worked for the past decade on various aspects of AIDS research.

Martin Delaney

Project Inform,

98 Cazneau,

Sausalito, California 94965, USA

SIR — While it is hard to disagree with Fields' call for funding of basic research in the fight against AIDS, his proposal may result in throwing out the baby with the bath water. Leaving aside the difficult question of what distinguishes 'basic' from 'applied' HIV research, bench scientists studying the interactions of HIV and the immune system have found it invaluable to have access to human samples from experimental HIV vaccine volunteers. These specimens have facilitated investigations of the fine specificity of cytotoxic T cell responses, antigen and virus induced production of Th1 and Th2 cytokines, proliferative responses of cells to various strains of HIV envelope following immunization and the generation of antibody responses (including the phenomena of cross-reactivity, cross-neutralization and "original antigenic sin").

Precise knowledge of the timing of exposure, genotype and phenotype of isolated gene products, and baseline immune status of vaccine volunteers, is particularly useful in designing controlled experiments. This knowledge is a luxury not usually afforded scientists when studying infected individuals (even prospectively). In short, if one were designing a national programme for the basic study of immune system HIV interactions and did not already have an experimental HIV vaccine network in place, it would have to be created.

It should also be noted that private industry has probably invested at least as much as the US government in HIV vaccine development which, like most vaccine development, is a partnership between industry, government and academic institutions. The concerns of private-sector partners should at least be considered when policy is formulated to reconfigure or eliminate government programmes that depend on private sector participation.

David H. Schwartz

Department of Immunology and

Infectious Diseases,

Johns Hopkins University,

624 N. Broadway,

Baltimore, Maryland 21205-1901, USA

Localizing electrons in atomic orbits

A simple technique for preparing atoms in such a way that a distant electron behaves for several periods of revolution as if it were a particle may be the most vivid illustration yet of the Correspondence Principle.

DECADES have passed since people pictured electrons orbiting around atoms as if they were point electrical charges moving in keplerian ellipses around the central nucleus. Bohr's first attempt at an understanding of atomic structure in 1913 was, after all, nothing more than an account of the solution of a problem in classical mechanics constrained by the empirically derived (or guessed) quantum conditions. A decade later, with the Uncertainty Principle (but the same quantum conditions), people took to heart the doctrine that an electron in an atom is a probability distribution of charge. Keplerian orbits were gone for good.

What follows is a piece of fun, a demonstration that the old concept of semi-classical Bohr orbits may not be entirely untenable. Those concerned are Z. D. Gaeta, Michael W. Noel and D. R. Stroud from the University of Rochester, New York. What they have done is to suggest how it would be possible to prepare the state of a hydrogen atom in such a way that the single electron would be seen (by suitably fast laser pulses) to be moving in a kind of orbit. The argument appears in what must be a vintage issue of *Physical Review Letters* (73, 636; 1 August 1994).

For what it's worth, the objectives underlying this kind of trickery are not entirely the maximization of collective amusement. It is at least conceivable that the controlled preparation of the electronic states of molecules by well-judged laser excitation may make particular chemical reactions more feasible than they would otherwise be. But that prospect is still some way off.

What Gaeta and his colleagues are for the time being concerned with is somehow to arrange that wave packets of electron density will move unchanged in arbitrary elliptical orbits around a hydrogen nucleus. They expect that they will be able to do this only if the electrons are relatively far from the nucleus, meaning that they will have a principal quantum number, n , that is relatively large, or that the atom is in a Rydberg state. That is the semi-classical regime, where electrons behave more or less as if they were classical entities. Then it is possible to use classical arguments to reach some kind of explanation of how the trick may be accomplished.

Because quantum theory is an exact description and classical mechanics only an approximation in restricted conditions, it is important to be clear what is going on in

these experiments (which can be done in the laboratory as well as the head). The s states of electrons around hydrogen atoms are, of course, represented by spherically symmetrical wave-functions whose algebraic form is well known. So why not take groups of them together in linear combinations to form atoms in which the electron distribution is a spherical annulus? There is nothing in principle to prevent that being done.

It is even possible to go further and, by taking linear combinations of orbitals that have no projection of angular momentum in some arbitrary direction, to construct orbitals (not pure states) that are ring-shaped. But this is not mere hand-waving. It is possible to prepare atoms in just those states in the laboratory, for example by using a static electric field and a laser tuned to the energy of the ring-like state. Inevitably, of course, there are limits (determined by the Uncertainty Principle) on the narrowness of such a ring; the greater the quantum number n , the relatively narrower the ring can be.

Gaeta and his colleagues first set out to show that it is possible to turn a ring-like Rydberg state into one in which there is a bunch of electron density moving around a circular orbit. Their recipe simply entails subjecting these atoms to a brief but relatively strong electric pulse for a picosecond or less. Because the ring-like states are already semi-classical, it is possible to use classical arguments to understand why this works.

Suppose, for example, that there are several (classical) electrons in the same circular orbit about a nucleus, and then arrange that there is a uniform electric field in the plane of the orbit for a brief interval of time, a small fraction of the period of the electrons in their orbit. Then particles instantaneously in the direction of the field will be accelerated, those moving in the opposite direction will be decelerated and those moving transversely to the field will not be affected.

The result of that velocity dispersion is that, eventually, all the electrons will bunch together briefly at some point in the orbit. So brief pulses of electric field are the means by which electrons are localized in circular orbits. Gaeta and colleagues show these bunches of electron density (each representing a single electron) persisting for several orbital periods, eventually decaying into gaussian smudges which then become uniform distributions about the circular orbit again. That the bunches of electron distribu-

tion last for even a single period is, of course, the surprise in this observation.

So why not go further, and make elliptical orbits by the same techniques? This is what proves possible. First, it appears, the ring-like orbits can be simply made into elliptical orbits by a steady electric field, whereupon the electrons can be made to travel around them by means of a short sharp field, as in the preparation of bunched circular orbits. The motion of the electrons has even been observed (with the help of femtosecond lasers).

As Kepler (or at least Heisenberg) might have predicted, an electron moves most quickly, and is less tightly bunched together, when near the nucleus, and is most tightly localized and most slowly moving at the other end of the orbit.

The immediate importance of these developments is, for the time being, pedagogical. They illustrate as vividly as anybody could ask the importance of the Correspondence Principle by which Bohr first grounded his semi-classical theory on observation, and which has since become the means of telling when a system has to be dealt with by quantum rather than classical mechanics. Here one can see electrons in that half-way stage between classical and quantum behaviour, making several orbits around a nucleus in a localized condition before eventually smearing out into a uniform distribution around the orbit.

Beyond that, the localized electron in orbit is a telling measure of how much has been learned of Rydberg atoms in the past decade or so. It is not merely that they have been powerful tests of atomic theory, but they have also been the proving-ground of the interaction of magnetic fields with atoms. In reality, for technical reasons and especially the lack of easily usable and tunable lasers in the ultraviolet, most studies have been made with Rydberg-like atoms, in which one of several electrons in a multi-electron atom has been excited into a state with large quantum number. But that may yet change.

Meanwhile, the fun and games mostly lie ahead, with the studies yet to come of the behaviour of Rydberg-like molecules. That is when it may be possible to learn something of the degree to which simple exposure to electric fields may bring about chemical reactions at present only possible in much more extreme conditions.

John Maddox

Our primary source of ATP

Richard L. Cross

A HIGH-RESOLUTION structure for bovine mitochondrial F_1 -ATPase is presented on page 621 of this issue by Abrahams *et al.*¹, which will be exciting news not only for the hundreds of laboratories that have worked on this important enzyme during the 34 years since it was first isolated², but also for structural biologists generally be-

only capable of net ATP hydrolysis, hence it is referred to as an ATPase.

A model for the *E. coli* synthase⁴, $\alpha_3\beta_3\gamma\delta\epsilon ab_2c_{9-12}$, is shown in Fig. 1. An outer cylindrical complex of $\alpha_3\beta_3c_{9-12}$ (blue/green) is depicted as surrounding an asymmetric core of $\gamma\delta\epsilon ab_2$ (yellow/red).

As stressed by Abrahams *et al.*¹, their

cycle; and (3) the binding changes required for catalysis are driven by the rotation of an asymmetric aggregate of subunits (like those shaded in yellow/red in Fig. 1) relative to the remaining subunits. Like a truncated bacterial flagellar motor, this rotation would be driven by protons moving down an electrochemical gradient.

Although features (1) and (2) are supported by a wealth of data, the rotational feature remains speculative. It was originally proposed as a means of accommodating the known asymmetric structure of F_1 with the observation of a single reaction pathway for the multiple catalytic sites⁵. Nature may have devised other ways to accomplish the necessary binding changes, however, and additional experiments will be required to settle this issue.

The structure presented by Abrahams *et al.* sheds light on several important mechanistic issues. Reflecting the strong cooperative interactions that are an integral part of the binding-change mechanism, the enzyme has crystallized in a form that has each catalytic site in a different conformation. Comparison of the structure of the open, empty catalytic site with that of a closed, filled site provides an explanation in terms of the movement of specific structural elements for the conformational change that deforms the site during product release. The distinct interactions between the single γ -subunit and the three different conformational states of the catalytic subunit are particularly interesting, and are consistent with the idea that an asymmetric inner core helps to impose asymmetry on the three catalytic subunits. Finally, it turns out that a glutamic acid is properly positioned at the catalytic site to activate a water molecule for attack on the γ -phosphoryl group of ATP, and an arginine is available to stabilize the negative charge on the resulting phosphate. The absence of any equivalent carboxylate at the non-catalytic site may explain why this site lacks hydrolytic activity.

The high-resolution structure also

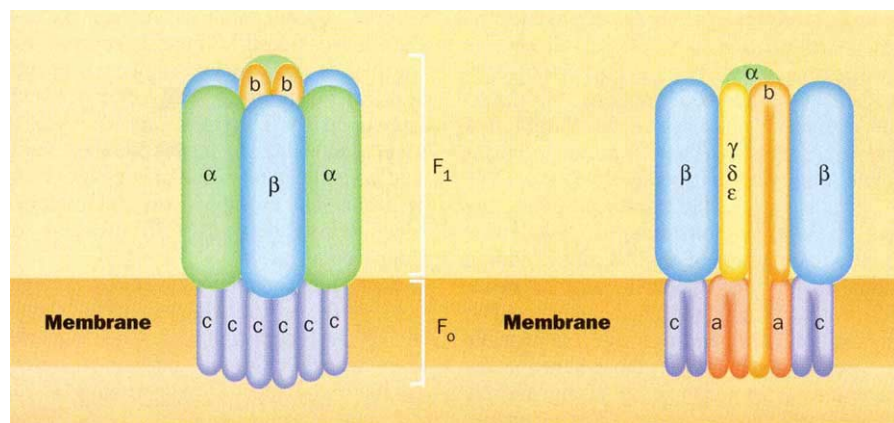


FIG. 1 Model for the structure of *E. coli* F_0F_1 . Left, complete complex; right, complex cut in half.

cause this is the largest asymmetric structure ever to be solved. And for those interested in the features and evolution of nucleotide-binding sites, of which F_1 has six, catalytic and non-catalytic, with strong cooperative interactions between them, this structure should prove a goldmine of information.

F_1 -ATPase, normally found in an F_0F_1 complex in the membranes of eubacteria, mitochondria and chloroplasts, is responsible for the synthesis of ATP by oxidative and photo-phosphorylation. An electrochemical H^+ gradient appears to serve as an obligatory intermediate in coupling the exergonic oxidation/reduction reactions of electron-transport chains to the endergonic synthesis of ATP by the synthase³. Close cousins of the standard F_0F_1 complexes are found as ATP synthases in archaeobacteria and as H^+ -translocating ATPases in the vacuolar membranes of eukaryotic cells. Nature makes frequent use of complexes of the F_0F_1 type, and these play a vital role in maintaining cellular ATP levels and in driving secondary transport systems.

The study of the reaction steps catalysed by the synthase has been aided by the separation of F_0F_1 into a soluble hydrophilic portion, F_1 , containing five subunits (α , β , γ , δ and ϵ), and a membrane-embedded hydrophobic part, F_0 , containing a minimum of three subunits (a , b and c in *Escherichia coli*). F_0 functions as a proton channel and F_1 contains the catalytic sites for ATP synthesis. Uncoupled from its energy source, solubilized F_1 is

results support the binding-change mechanism that, over the past two decades, has become widely used in describing ATP synthesis by F_0F_1 (refs 5, 6). As illustrated in Fig. 2, the three main tenets of the hypothesis are: (1) the primary use of energy supplied by proton flow through F_0 is not to make ATP, but rather to promote its release from catalytic sites where it forms spontaneously; this reflects the need to disrupt the many favourable interactions between protein and ligand that contribute to the very tight binding of ATP; (2) energy-linked substrate binding, formation of tightly bound ATP and product release occur simultaneously at three separate but interacting catalytic sites that remain 120° out of phase in the catalytic

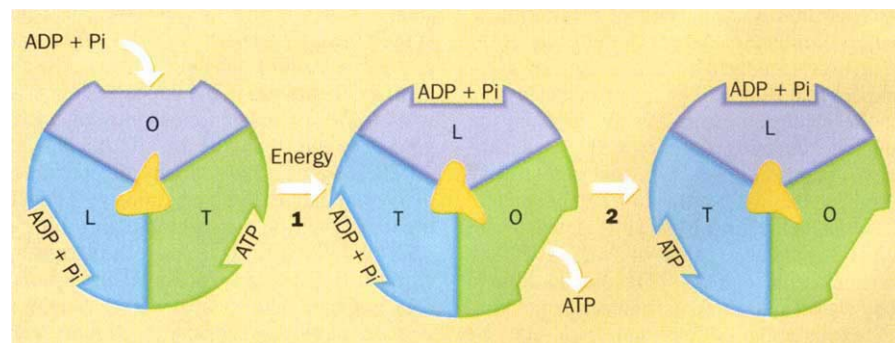


FIG. 2 The binding-change mechanism for ATP synthesis. The central asymmetric mass (yellow), representing $\gamma\delta\epsilon b_2$, rotates in step 1 relative to the three $\alpha\beta$ pairs (blue/green) which, in this illustration, remain stationary. This rotation forces the three catalytic sites to undergo conformational changes associated with substrate binding and product release. T, L and O stand for tight, loose and open conformations and refer to the affinities of catalytic sites for ligand. In step 2, ATP forms spontaneously from tightly bound ADP and P_i .

sheds light on earlier structural studies and their interpretations, confirming some and showing others to be incorrect. The structure of the nucleotide-binding sites confirms the presence of several residues previously identified by affinity labelling⁷⁻¹⁰. The positioning of the sites at α/β interfaces is substantiated^{11,12}, as is the importance of α -subunits in forming the non-catalytic sites^{13,14}. The arrangement of a hexameric aggregate of alternating α - and β -subunits surrounding the γ -subunit derived from immunoelectron-microscopy is also confirmed^{15,16}. However, proposals are not supported that the non-catalytic and catalytic sites are situated at the same α/β -subunit interfaces and oriented either like the sites on adenylate kinase¹⁷ or with the adenine moieties in close proximity¹². The suggestion that the terminal phosphoryl group of non-catalytic-site nucleotide acts as an acid/base catalyst for ATP formation at an adjacent catalytic site¹⁸ is also untenable.

This salient contribution from Abrahams *et al.*¹ caps a series of important earlier contributions from the Walker laboratory. They determined the sequence of the *unc* operon which codes for all subunits of *E. coli* F_0F_1 . Comparison of the sequences of the α - and β -subunits with those of other proteins led to the identification of homologies shared by many nucleotide-binding proteins, including the glycine-rich P-loop (also known as Walker's sequence A). Having sequenced bovine F_1 subunits and identified and sequenced subunits in bovine F_0 , they have now produced a high-resolution structure for F_1 , results that add up to a remarkable accomplishment.

The most immediate application of this new information will probably be to studies of the mechanism by mutagenesis.

With an accurate structure to guide experiments, progress in working out remaining details of this important energy-coupling system should be accelerated. The next dramatic advance in the field may be the solution of a high-resolution structure for F_0 . F_0 structures differing from that shown in Fig. 1 have been proposed, and F_0 and F_1 may not be in direct contact but separated by a stalk. Once the structure of F_0 is known, it should help clarify the possible role of subunit rotation. Also, when available,

comparison of the structure of F_1 from other sources¹⁹ with that of the bovine enzyme should provide additional insight. Indeed, F_0F_1 stands a good chance of being the first ATP-driven membrane transport system to be understood at the molecular level. □

Richard L. Cross is in the Department of Biochemistry and Molecular Biology, State University of New York Health Science Center at Syracuse, Syracuse, New York 13210, USA.

LASER PHYSICS

Hard X-rays from atom clusters

Bernd Crasemann

THE elusive aim of achieving a readily usable laser which emits short-wavelength X-rays may be coming closer to our grasp with the discovery of a new mechanism for multiphoton excitation of atomic inner shells in rare-gas clusters¹. On page 631 of this issue, members of Charles Rhodes' group report measurements of 2–3-Å (4–5-keV) X-rays from xenon clusters irradiated with short pulses of very intense ultraviolet light². In conjunction with results from previous work on krypton clusters³, the data provide evidence for a strong mode of electromagnetic coupling through coherent multi-electron motion induced by the intense external field; in this process, photoelectrons produce copious core-excited ions that promptly emit energetic X-rays. An improved scaling relationship is attained between excitation power and gain.

Efforts to design shorter-wavelength lasers ensued soon after it became apparent, in the late 1950s, that devices using amplification by stimulated emission could be obtained in the near-visible spectral region⁴. A big step forward was the demonstration in 1985 of strong amplification of spontaneous emission from a 200-Å transition in neon-like selenium ions, driven by the large NOVA laser in the Lawrence Livermore National Laboratory⁵. Considerable progress ensued in the soft X-ray regime⁶. The road to shorter wavelengths has, however, turned out to be tortuous. Groups in many parts of the world are at present working on X-ray lasers. This large effort is motivated by the potential of such devices for a wide range of applications, from biological imaging to uses in atomic and molecular physics, materials science and chemistry, as well as industrial processes and defence projects.

The crux of the difficulty in reaching shorter laser wavelengths is the escalating requirement for higher pumping power to achieve population inversion in the

medium⁷. Several factors contribute to this effect. The lifetime of relevant atomic excited states is roughly proportional to λ^2 , the square of the wavelength, so that the number of atoms that must be pumped up per unit time is proportional to λ^{-2} . The transition energy, ϵ , being proportional to λ^{-1} , it follows that the energy per unit time required to maintain the excited population goes as λ^{-3} . For example, to maintain a vacancy (with a lifetime of 1 femtosecond) in the innermost shell of one copper atom, an 8-keV photon (1.5-Å wavelength) needs to be absorbed 10^{15} times per second: it takes about 1 watt to maintain this K-shell vacancy. The entire power consumption of the United States would suffice only to keep 0.1 nanogram of copper pumped up continuously. Furthermore, the laser amplification constant which is proportional to the cross-section for stimulated emission scales as λ^{-2} so that the number of simultaneously excited atoms must be raised accordingly. This crude analysis shows that the pumping power required for a given amplification must be roughly proportional to λ^{-5} , the inverse fifth power of the wavelength.

The obvious recourse is to aim at achieving population inversion only for exceedingly short periods of time and to search for the most effective means of delivering energy to a restricted volume of the medium. Concentration of power can thus be seen as the central problem in X-ray laser design. A solution can be attempted by seeking new modes of energy delivery and concentration that establish a more favourable scaling relationship between the magnitude of gain produced and excitation power required for X-ray amplification.

And this is the key point of the work reported here by McPherson *et al.*². Two physical phenomena are assumed to play a role in producing large power compression in these experiments: a novel mode of multiphoton excitation of clusters and a relativistic regime of high-intensity pulse

1. Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. *Nature* **370**, 621–628 (1994).
2. Penefsky, H. S., Pullman, M. E., Datta, A. & Racker, E. J. *biol. Chem.* **235**, 3330–3336 (1960).
3. Mitchell, P. *Nature* **191**, 144–148 (1961).
4. Foster, D. L. & Fillingame, R. H. *J. biol. Chem.* **257**, 2009–2015 (1982).
5. Boyer, P. D. *Biochim. biophys. Acta* **1140**, 215–250 (1993).
6. Penefsky, H. S. & Cross, R. L. *Adv. Enzym. rel. Areas molec. Biol.* **64**, 173–214 (1991).
7. Garin, J., Boulay, F., Issartel, J. P., Lunardi, J. & Vignais, P. V. *Biochemistry* **25**, 4431–4437 (1986).
8. Cross, R. L. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 5715–5719 (1987).
9. Bullough, D. A. & Allison, W. S. *J. biol. Chem.* **261**, 14171–14177 (1986).
10. Tagaya, M., Nouri, T., Nakano, K., Futai, M. & Fukui, T. *FEBS Lett.* **233**, 347–351 (1988).
11. Schafer, H.-J. & Dose, K. *J. biol. Chem.* **259**, 15301–15306 (1984).
12. Zhuo, S., Garrod, S., Miller, P. & Allison, W. S. *J. biol. Chem.* **267**, 12916–12927 (1992).
13. Dunn, S. D. *J. biol. Chem.* **255**, 11857–11860 (1980).
14. Maggio, M. B., Pagan, J., Parsonage, D., Hatch, L. & Senior, A. E. *J. biol. Chem.* **262**, 8981–8984 (1987).
15. Tiedge, H. & Schafer, G. *Biochim. biophys. Acta* **977**, 1–9 (1989).
16. Capaldi, R. A., Aggeler, R., Turina, P. & Wilkens, S. *Trends biochem. Sci.* **19**, 284–289 (1994).
17. Vogel, P. D. & Cross, R. L. *J. biol. Chem.* **266**, 6101–6105 (1991).
18. Harris, D. A. *FEBS Lett.* **316**, 209–215 (1993).
19. Blanchet, M., Ysem, X., Hüllihen, J., Pedersen, P. L. & Amzel, L. M. *J. biol. Chem.* **266**, 21197–21201 (1991).

propagation in plasmas.

In atoms that make up a cluster, the intense external field induces oscillations of the outer, weakly bound electrons. This motion of outer electrons is coherent, constituting an oscillating charge cloud; it can lead to very efficient energy transfer to tightly bound inner-shell electrons⁸, provided that certain relationships between the intensity of irradiation and the cluster size are satisfied¹. The energy of numerous pumping photons can thus be transferred via photoelectrons from outer shells to cause ionization of an inner-shell electron, producing a deep atomic vacancy that is subsequently filled under prompt emission of a hard X-ray. The evidence even points towards multiple inner-shell ionization, producing 'hollow atoms'. This process thus leads to the delivery of high local power density.

The second phenomenon is indicated by evidence for the formation of a stable mode of spatially confined (channelled) propagation of the intense subpicosecond ultraviolet radiation in the plasma⁹. This self-channelling arises from the combined action of the relativistic mass shift and displacement of the electron component of the plasma. It leads to the required spatial organization in the delivery of pumping power, producing multiphoton inner-shell electron ejection along channels of ionization. In a discussion of his findings, Rhodes points out that both of these new phenomena arise because the coherent multi-electron motion induced by the external field simulates the behaviour of a light particle carrying a high charge, thereby transforming the fundamental coupling strength of the interaction of electrons with matter from the customary factor α ($\approx 1/137$) to the enhanced value $N^2\alpha$, where N is the number of electrons participating in the coherent multi-electron amplitude of oscillation in an atom.

This scheme of multiphoton excitation of clusters and confined modes of electromagnetic propagation gives us the hope of achieving exceptional phase-space control over the deposited energy. It could lead to both the necessary power density and to confinement for X-ray amplification well in excess of other known approaches. □

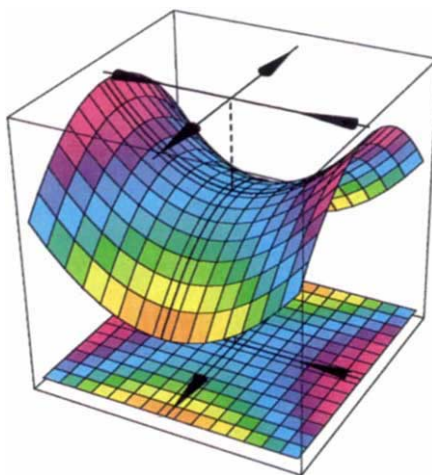
Bernd Crasemann is in the Physics Department, University of Oregon, Eugene, Oregon 97403, USA.

1. McPherson, A. *et al.* *Appl. Phys.* **B57**, 338–347 (1993).
2. McPherson, A., Thompson, B. D., Borisov, A. B., Boyer, K. & Rhodes, C. K. *Nature* **370**, 631–634 (1994).
3. McPherson, A. *et al.* *Phys. Rev. Lett.* **72**, 1810–1813 (1994).
4. Schawlow, A. L. *IEEE Quantum Electron.* **QE9**, 646–647 (1973).
5. Matthews, D. L. *et al.* *Phys. Rev. Lett.* **54**, 110–113 (1985).
6. Key, M. H. *Mod. Opt.* **35**, 575–585 (1988).
7. Elton, R. C. *X-ray Lasers* (Academic, New York, 1990).
8. Boyer, K. & Rhodes, C. K. *Phys. Rev. Lett.* **54**, 1490–1493 (1985).
9. Borisov, A. B. *et al.* *Phys. Rev. Lett.* **68**, 2309 (1992).

Chaos under control

Frank Moss

To control chaos in living systems — indeed, even to detect and demonstrate it reliably — holds out the rosy promise of entirely new therapeutic and diagnostic tools for diseases ranging from heart disease to epilepsy. On page 615 of this issue¹, Steven Schiff, Mark Spano, William Ditto and their colleagues now show that a neural network prepared from a hippocampal slice of rat brain not only can be made to demonstrate chaotic be-



A saddle instability. The unstable periodic point is located at the intersection of the surface with the dashed line. The stable and unstable directions are respectively indicated by the inward- and outward-pointing arrows on the top and bottom face projections.

haviour, but that the chaos can be controlled. Moreover, periodic behaviour in certain preparations can be *anticon*trolled in order to induce chaos.

Studies on chaos in physico-chemical systems have a well-established history, and mathematical algorithms for calculating the perturbations necessary to control chaotic systems (that is, to render their behaviour regular or periodic) have recently emerged². These algorithms have spawned a minor industry of sorts, so that chaotic magneto-elastic devices³, lasers⁴ and even chemical reactions⁵ have been controlled. Now add biological systems to the list — perhaps the most elusive but potentially beneficial of all.

The brain slice, taken from the temporal lobe, is deliberately driven into chaos by immersing it in an artificial K^+ -enriched cerebrospinal fluid, whereupon synchronous bursts originating in the cornu ammonis propagate along a recurrent fibre tract to points where they are recorded. The interburst interval times, T , which exhibit considerable variability, are recorded on a first-return scatter plot (T_n versus T_{n-1}). Sauer has recently shown

how to analyse such intervals, generated by the threshold crossings of an underlying chaotic process⁶. A perfectly periodic (period one) interburst interval appears as a single dot at $T_n = T_{n-1}$ lying on the diagonal of the scatter plot. One of the characteristic features of chaotic attractors is that they are built upon sets of unstable periodic points. Every such period-one point lies on the diagonal, but because they are unstable the points are not visited often. Typically, a trajectory consists of a sequence of points approaching the neighbourhood of the unstable point along a stable direction with ever-decreasing distances from the diagonal, and departing this encounter along a different, unstable, direction with ever-increasing point-to-diagonal distances. And these encounters are recurrent. In the jargon of dynamical systems theorists, such a point is called a 'saddle' and the directions are the stable and unstable 'manifolds' (see figure).

To observe such events rigorously is to observe chaos. The chaos can then be actively controlled by perturbing the system (for instance, by precisely timed electrical stimulation of a brain slice) so that each would-be errant trajectory is targeted to land precisely on the stable direction in the neighbourhood of the periodic point. The chaotic dynamics then pulls the system towards the point, and the ensuing trajectories become periodic. Some periodic points are naturally stable. Anticontrol is accomplished by perturbing a naturally stable trajectory so that it will fall well away from the neighbourhood of the periodic point.

The thinking behind such manipulations is that they may point to possible therapies for cardiac disease and even epilepsy, because it has been suggested that the healthy state of the heart or brain is chaotic, and that diseases of these organs are often presaged by more regular, or even periodic, behaviour⁷. This

1. Schiff, S. J. *et al.* *Nature* **370**, 615–620 (1994).
2. Shinbrot, T., Grebogi, C., Ott, E. & Yorke, J. A. *Nature* **363**, 411–417 (1993).
3. Ditto, W. L., Raue, S. N. & Spano, M. L. *Phys. Rev. Lett.* **65**, 3211–3214 (1990).
4. Roy, R., Murphy, T. W., Maier, T. D. & Gills, Z. *Phys. Rev. Lett.* **68**, 1259–1262 (1992).
5. Petrov, V., Gaspar, V., Masere, J. & Showalter, K. *Nature* **361**, 240–243 (1993).
6. Sauer, T. *Phys. Rev. Lett.* **72**, 3811 (1994).
7. Goldberger, A. L., Rigney, D. R. & West, B. J. *Sci. Am.* **262**, 42–49 (1990).
8. Chang, T. *et al.* *Biophys. J.* **87**, 671–683 (1994).
9. Ruelle, D. *Physics Today* **47**, 24–30 (1994).
10. Grassberger, P. & Procaccia, I. *Physica D* **9**, 189–208 (1983).
11. Kaplan, D. T. & Glass, L. *Phys. Rev. Lett.* **68**, 427–430 (1992).
12. Cvitanović, P. *Phys. Rev. Lett.* **61**, 2729–2732 (1988).
13. Garfinkel, A., Spano, M. L., Ditto, W. L. & Weiss, J. N. *Science* **257**, 1230–1235 (1992).

pretty dynamical picture is unfortunately obscured in real systems, and certainly in biological preparations, by noise, or random behaviour not meeting the criteria that define the aforementioned encounter. The trick is convincingly to identify the chaotic encounters, and hence to locate the unstable periodic points, in a sea of noise. Having done that, both control and anticontrol, and maybe new diagnostic and therapeutic technologies, become feasible. Because of the great potential benefit, the search for chaos in biological and medical settings has been energetically pursued⁷; it has not yet been convincingly demonstrated, although some results are encouraging⁸.

Earlier this year, Ruelle published an excellent article outlining the obstacles⁹. There are at least three: data sets are too small; preparations, or patients, are not stationary; and the most commonly used detection algorithms^{10,11}, which require large data sets and hence also stationary

preparations, are susceptible to noise contamination. What Schiff *et al.* have done, for the first time in real-world experiments, is to make use of the fact that chaos is built upon an underlying structure — a skeleton — of unstable periodic points, or cycles, as Cvitanović first called them¹². In contrast to the more commonly used algorithms, the present method shows a degree of noise immunity and can be used with comparatively short records. Consequently, the detection of recurrent, unstable cycles, and thus of chaos, in a statistically convincing way might just become routinely possible with biological preparations or even with patients. These new experiments¹, together with a previous one by a subset of the same authors¹³, open the door to this grand possibility. □

Frank Moss is in the Department of Physics and Astronomy, University of Missouri at St Louis, St Louis, Missouri 63121, USA.

TRANSLATION

Dissecting RNA function

Renée Schroeder

TRANSLATION of a messenger RNA into protein requires the interaction of an aminoacylated transfer RNA with the cognate codon on the mRNA. The stability of Watson–Crick base pairing between codon and anticodon is not sufficient in itself to assure efficient and accurate decoding and needs assistance from the ribosome. Is this function achieved by ribosomal proteins, by ribosomal RNA or by a combined action of both? On page 659 of this issue, Purohit and Stern¹ provide evidence that it is the rRNA that mediates the interaction of mRNA with tRNA. A small RNA was designed to mimic the 16S rRNA decoding region in order to investigate its function as the key molecule in the recognition of the codon/anticodon interaction. This small oligoribonucleotide was shown by footprinting experiments to interact autonomously with aminoglycoside antibiotics and RNA ligands, as predicted for the decoding function. The dissection of large multifunctional RNAs into small autonomously active subdomains represents a potent strategy in the analysis of large RNA molecules.

In the past decade, attention has focused on rRNAs after a long period of belief that proteins were the functional elements in translation. The peptidyl transferase activity of the large ribosomal subunit is surprisingly resistant to protein extraction procedures, indicating that the 23S rRNA plays a key role in peptide bond synthesis². Purohit and Stern's experiments now suggest that decoding is an

RNA function. The decoding oligoribonucleotide was designed to comprise the peptidyl tRNA (P) and aminoacyl tRNA (A) binding sites of the 16S rRNA and was tested for its ability to bind aminoglycoside antibiotics, tRNAs and mRNA. Remarkably, the P-site of the subdomain analogue interacted selectively with RNA ligands, discriminating between anticodon stem-loops and tetraloops and another smaller stem-loop structure containing a three-U loop. The A-site subdomain interacted with both tRNA and mRNA alone (poly(U) and poly(C)), whereas the ribosomal A-site in the small subunit is strongly dependent on the presence of mRNA for tRNA binding³. Although these differences were detected between the activities of the 30S ribosomal subunit and the decoding analogue, the results clearly show that binding of aminoglycoside antibiotics that induce miscoding, of mRNA, and of a tRNA anticodon stem-loop is an intrinsic property of the decoding region of the 16S rRNA. But it remains to be demonstrated that this small decoding-subdomain analogue recognizes the codon/anticodon complex.

Interaction of the codon–anticodon duplex with the decoding region of 16S rRNA is only the first step in the decoding process. The aminoacyl tRNA is delivered to the ribosome as a ternary complex with the elongation factor EF-Tu bound to GTP. Once the correct tRNA is bound, a signal has to be transmitted for a peptide bond to be formed. It has recently been

suggested⁴ that the ribosome senses the binding of the correct tRNA and transmits this information through a conformational change in the 16S rRNA to trigger GTP hydrolysis and release of EF-Tu. Several ribosomal proteins (S4, S5 and S12) and particularly the RNA domains surrounding the G530 loop and the 912–915 region (see Fig. 1a in ref. 1) seem to be involved in this signalling step. Whereas recognition of the codon/anticodon complex might be an evolutionarily early function of the RNA, translation accuracy can be viewed as an evolutionarily late RNA/protein function representing a refinement of the pre-existing RNA-mediated decoding mechanism⁵.

The decoding function might have a counterpart in self-splicing group I introns. Selection of the splice site for initiation of the first step of splicing involves recognition of a short RNA duplex by the catalytic core of the intron in a manner that resembles decoding^{6,7}. Like the decoding function, this process is perturbed by aminoglycoside antibiotics^{8,9}. Purohit and Stern's experiment has a significant impact on the evolutionary picture of the ribosome. It is difficult to imagine how the ribosome and even rRNAs evolved, owing to their size and complexity. The experiments presented here support the idea that small autonomously acting RNAs were at the origin of translation. RNA molecules have two intrinsic properties that make them optimal as evolutionarily early molecules. First, they fold into active structures even when very small, so that catalytically active sites exist on their own, and second, these small autonomous RNAs are able to interact with each other in *trans*. Examples of such *trans*-active RNAs are found in self-splicing group II introns, which can be complemented with their essential domains in *trans*¹⁰, in the spliceosome, which is an RNA/protein complex whose RNA components interact in *trans*, and in some ribosomes that are assembled from scrambled rRNA molecules¹¹. These intrinsic properties of RNA enable large multifunctional RNAs to be dissected into small autonomously active subdomains, as demonstrated by Purohit and Stern¹,

1. Purohit, P. & Stern, S. *Nature* **370**, 659–662 (1994).
2. Noller, H. F., Hoffarth, V. & Zimniak, L. *Science* **256**, 1416–1419 (1992).
3. Moazed, D. & Noller, H. F. *Cell* **46**, 985–994 (1986).
4. Powers, T. & Noller, H. F. *J. molec. Biol.* **235**, 147–172 (1994).
5. Noller, H. F. in *The RNA World* (eds Gesteland, R. F. & Atkins, J. F.) 137–156 (Cold Spring Harbor Laboratory Press, New York, 1993).
6. Pyle, A. M., Murphy, F. L. & Cech, T. R. *Nature* **358**, 123–128 (1992).
7. von Ahlsen, U. & Noller, H. F. *Science* **260**, 1500–1503 (1993).
8. von Ahlsen, U., Davies, J. & Schroeder, R. *Nature* **353**, 368–370 (1991).
9. Schroeder, R., Streicher, B. & Wank, H. *Science* **260**, 1443–1444 (1993).
10. Jarrell, K. A., Dietrich, R. C. & Perlman, P. S. *Molec. cell. Biol.* **8**, 2361–2366 (1988).
11. Boer, P. H. & Gray, M. W. *Cell* **55**, 399–411 (1988).

which can then be subjected to extensive analysis, usually impossible on large molecules directly.

The 'RNA world' hypothesis is gaining credibility as more RNA-governed activities are revealed. The discovery that peptidyl transferase activity might be an RNA-mediated reaction was unexpected; now decoding is joining the RNA 'club'. Are we soon going to be surprised by finding a ribosomal protein that has an essential function? Today's ribosome

evolved to function with rRNAs, proteins and translation factors acting together to ensure translational efficiency and accuracy. Pains-taking investigation of the ribosome's different functions will help us understand the complex and subtle interplay between the components of the translational apparatus. □

Renée Schroeder is in the Institute of Microbiology and Genetics, University of Vienna, Dr Bohrgasse 9, A-1030 Vienna, Austria.

HIGH-TEMPERATURE SUPERCONDUCTORS

The search for symmetry

Victor J. Emery

AT a conference on superconductivity held in July*, the liveliest discussion centred on the symmetry of the ordered state of high-temperature superconductors — so lively in fact that the debate spilled out into the corridors, and an impromptu session had to be organized to give everyone a chance to air their views. This apparently esoteric property is important for what it may reveal about the nature of high-temperature superconductivity, which has sparked so much effort around the world since the initial discovery in 1986. Knowledge of the symmetry in question would not unequivocally identify the underlying mechanism of high-temperature superconductivity but it could serve to narrow the field.

The simplest way to think about the problem is to focus on the copper oxide planes in which the primary superconducting charges reside, and to imagine that the copper ions form a square lattice. There are good reasons to believe that the superconducting state of the cuprates, like that of conventional superconductors, is composed of pairs of electrons with even angular momentum. For a square lattice the allowed quantized even angular momenta are zero (*s*-wave) and two (*d*-wave). Typically a short-range attractive interaction gives rise to *s*-wave pairing but a short-range repulsion combined with a longer-range attraction may favour a higher angular momentum state, which serves to keep the colliding particles away from each other and optimize the attraction. This is the essence of the argument that led to suggestions of *d*-wave pairing in liquid ^3He in 1960.

From this perspective, the nature of the pairing state helps us to identify the form of the effective interaction between the conduction electrons. But that is not the whole story: collective effects such as spin fluctuations may come into play and either

modify or completely determine the nature of the pairing state. In the case of liquid ^3He , ferromagnetic spin fluctuations lead to *p*-wave pairing, with odd angular momentum. About ten years ago, it was suggested that the exchange of antiferromagnetic spin fluctuations might be the source of superconductivity in organic and heavy fermion materials, and it was naturally one of the many mechanisms considered in the first frantic days of high-temperature superconductivity, especially after the discovery that the 'parent' insulating compounds are antiferromagnetically ordered.

Since then the idea has been pursued vigorously by several groups around the world. This is the mechanism that raises the stakes in the search for symmetry, for, at least in its simplest form, it is ruled out if the pairing is *s*-wave. Antiferromagnetic spin fluctuations carry one unit of spin and, as they are emitted or absorbed, they effectively flip the spin of a conduction electron. For singlet pairs, the spin flip brings along a sign change which converts an otherwise attractive exchange force into a repulsion. The saving grace is that the interaction follows the spatial oscillations of the spin configurations, so the pairs may arrange their angular momentum to catch the oscillations on the rise and experience an effective attraction. The lowest, and most effective, angular momentum state of this kind is *d*-wave.

What does experiment have to say? Much hard work, both theoretical and experimental, has gone into this question. For the most thoroughly studied material, $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$, there are clear indications that the energy gap is anisotropic and that it is zero at some places on the Fermi surface. But that can be true of either *s*-wave or *d*-wave. The experiments which are stirring so much interest in the community have addressed the question more directly by asking if the average value of the energy gap vanishes, or alternatively by looking for evidence that sign of the energy gap is changed by a 90° rotation,

both of which would be symptoms of *d*-wave pairing.

At this point the message is mixed. If we take the existing experiments at face value and thereby sidestep much of the current debate, it appears that the energy gap has both *d*-wave¹⁻³ and *s*-wave⁴⁻⁶ components.

On closer examination, this should not be surprising. Most discussions have proceeded on the assumption that, although $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ is an orthorhombic material, the simple square lattice analysis is not too bad. But this assumption is questionable: in $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ there is a significant electronic anisotropy in the form of a set of conducting CuO chains running along one axis of the copper oxide planes. The chains may participate in the superconductivity through their proximity to the CuO_2 planes and, in turn, impose an anisotropic potential onto the planes themselves. Strictly speaking, the different angular momentum states must be mixed, and there is no reason to believe that the mixing is insignificant. Indeed the superfluid density is quite different along the directions parallel to and perpendicular to the chains. All of this leaves us with the much more difficult task of deciding which is the dominant angular momentum state, and that requires a much more detailed theoretical analysis than has been carried out so far. A similar argument may be made for $\text{Bi}_2\text{SrCaCu}_2\text{O}_{8+x}$, which is anisotropic because of a lattice modulation that is known to influence the electronic properties, although to a smaller degree than the chains in $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$.

Obviously it would be cleaner to carry out the same analysis on tetragonal materials, for which the different angular momenta are not mixed. The problem appears to be that available $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ samples have been of much higher quality than their tetragonal cousins. Nevertheless, point contact tunnelling⁷ indicates that $\text{HgBa}_2\text{CuO}_{4+\delta}$ is *s*-wave, and there is convincing evidence⁸ of *s*-wave pairing in the electron-doped material $\text{Nd}_{2-x}\text{Ce}_x\text{CuO}_4$, which also shows striking antiferromagnetic behaviour. No one will be satisfied by such limited information, but experiments on tetragonal materials may well be the wave of the future. □

Victor J. Emery is in the Department of Physics, Building 510, Brookhaven National Laboratory, Upton, New York 11973, USA.

*Fourth International Conference on Materials and Mechanisms of Superconductivity/High-Temperature Superconductors, Grenoble, France, 5–9 July 1994.

1. Wollman, D. *et al.* *Phys. Rev. Lett.* **71**, 2134 (1993).
2. Tsuei, C. *et al.* *Phys. Rev. Lett.* **73**, 593 (1994).
3. Mathai, A., Gim, Y., Black, R. C., Amar, A. & Wellstood, F. C., *Univ. of Maryland preprint* (1994).
4. Chaudhari, P. & Lin, S. Y. *Phys. Rev. Lett.* **72**, 1084 (1994).
5. Sun, A. G. *et al.* *Phys. Rev. Lett.* **72**, 2267 (1994).
6. Buan, J. *et al.* *Phys. Rev. Lett.* **72**, 2632 (1994).
7. Chen, J. *et al.* *Phys. Rev.* **B49**, 3683 (1994).
8. Wu, D. H. *et al.* *Phys. Rev. Lett.* **70**, 85 (1993).

Ways of coping with stress

D. Grahame Hardie

BECAUSE humans usually have abundant reservoirs of glucose, and sophisticated homeostatic mechanisms for controlling them, blood glucose in a healthy adult scarcely drops even after weeks of total starvation. Death results from some associated complication, rather than starvation of cells for glucose itself. So it might seem curious that a system involved in the response to glucose starvation in the budding yeast *Saccharomyces cerevisiae* should have been highly conserved in mammals. Yet it is, as demonstrated in a paper by Carling and associates¹. This line of research has some way to go, but should offer a satisfyingly neat link between two previously unconnected physiological responses.

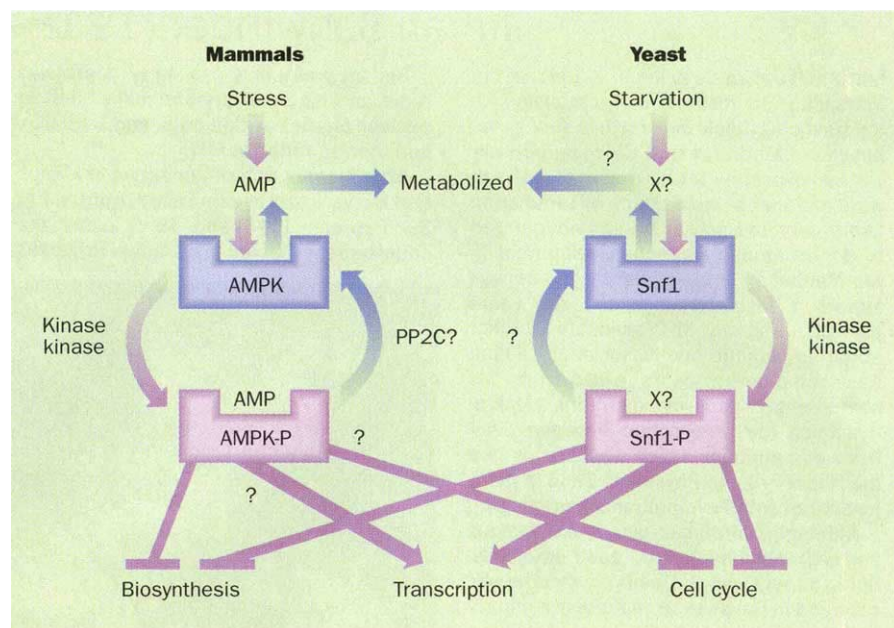
Mammalian AMP-activated protein kinase (AMPK) was identified as a protein that inactivates HMG-CoA reductase, the regulatory enzyme of isoprenoid biosynthesis. Subsequently it was shown to phosphorylate and inactivate a number of other biosynthetic enzymes, including acetyl-CoA carboxylase which regulates fatty-acid synthesis². AMP allosterically activates AMPK, but also promotes phosphorylation and activation by an upstream kinase kinase. The two effects of AMP multiply together, giving 100 times activation overall, and both effects are antagonized by high concentrations of ATP. This protein kinase cascade therefore represents a sensitive system which monitors the cellular AMP:ATP ratio. This ratio is normally very low in healthy cells, but stress treatments that inhibit oxidative phosphorylation, such as heat shock, hypoxia or arsenite, cause a dramatic rise in the ratio through the adenylate kinase reaction ($2\text{ADP} \leftrightarrow \text{ATP} + \text{AMP}$). AMPK is indeed switched on by cellular stress, and this correlates with inactivation of its target biosynthetic pathways, leading to the idea that it protects against stress by conserving ATP through inhibition of biosynthesis³.

Budding yeast prefers to grow by fermentation of glucose, and when glucose is abundant, genes required for oxidative metabolism and for use of other carbon sources (for example sucrose, galactose), are repressed; this is the phenomenon of carbon catabolite repression. Once glucose starts to run out, these genes can be derepressed, which absolutely requires the *SNF1* gene product. Cloning of *SNF1* in 1986 showed that it encodes a protein kinase⁴. Although no exogenous substrate was known at that time, Snf1p would autophosphorylate, and mutagenesis showed that the kinase activity was essential for function *in vivo*.

The next step was the cloning of the rat

complementary DNA encoding p63, the catalytic subunit of AMPK (ref. 5). It turned out to be closely related to Snf1p, a finding reached independently by the amino-acid sequencing of peptides derived from porcine p63 (ref. 6). Yeast Snf1p and rat AMPK have amino-terminal kinase domains that are some 60

genetic evidence it seems that it is also a substrate *in vivo*¹. The similarity also extends to regulation of the kinases. Like AMPK, Snf1p can be inactivated by protein phosphatases, and reactivated by mammalian kinase kinase, showing that the entire cascade has been conserved (see figure). Thus a system that appears to have evolved to protect against starvation (an everyday stress for yeast) may have been adapted as a response to other types of stress in mammals. These findings will allow the yeast system to



Known and putative parallels between the AMPK and Snf1p protein kinase cascades. Events occurring under stress/starvation conditions are shown in purple, and those occurring in unstressed cells in blue. Available evidence suggests that AMPK is dephosphorylated in intact cells by protein phosphatase-2C (PP2C) (ref. 2). Increase of AMP in response to stress activates AMPK by a dual mechanism: direct allosteric activation, and promotion of phosphorylation by the kinase kinase. The signal for starvation in yeast is unknown. Known downstream targets for AMPK, identified biochemically, are chiefly biosynthetic enzymes. Snf1p also phosphorylates and inactivates at least one biosynthetic enzyme (acetyl-CoA carboxylase) but genetic evidence suggests that it also has targets involved in regulation of transcription and the cell cycle.

per cent the same, but the similarity, although lower, also extends through the carboxy-terminal domains which may have a regulatory function.

It now emerges that this sequence relationship is reflected in functional similarities¹. AMPK is usually assayed using a synthetic peptide based on the sequence phosphorylated on rat acetyl-CoA carboxylase. Yeast extracts contain an activity that phosphorylates this peptide, and this is totally absent in a *snf1*-disrupted strain. Using this assay, Carling and colleagues show that Snf1p is activated by starvation, a point that was missed previously using the autophosphorylation assay. This effect is stable to purification (like the effect of stress on AMPK), suggesting that it arises from a covalent modification. Snf1p and AMPK also share at least one physiological substrate: Snf1p phosphorylates acetyl-CoA carboxylase *in vitro*⁶, although from

act as a model for the mammalian system and vice versa, yielding testable hypotheses.

The yeast model suggests that the mammalian kinase could also regulate gene expression, perhaps of stress-induced proteins. A number of yeast genes that seem to encode transcription

1. Woods, A. et al. *J. Biol. Chem.* **269**, 19509–19515 (1994).
2. Hardie, D. G. & MacKintosh, R. W. *BioEssays* **14**, 699–704 (1992).
3. Corton, J. M., Gillespie, J. G. & Hardie, D. G. *Curr. Biol.* **4**, 315–324 (1994).
4. Celenza, J. L. & Carlson, M. *Science* **233**, 1175–1180 (1986).
5. Carling, D. et al. *J. Biol. Chem.* **269**, 11442–11448 (1994).
6. Mitchellhill, K. I. et al. *J. Biol. Chem.* **269**, 2361–2364 (1994).
7. Lesage, P., Yang, X. & Carlson, M. *Nucleic Acids Res.* **22**, 597–603 (1994).
8. Thompson-Jaeger, S., Francois, J., Gaughran, J. P. & Tatchell, K. *Genetics* **129**, 697–706 (1991).
9. Wu, L. & Russell, P. *Nature* **363**, 738–741 (1993).
10. Parker, L. L., Walter, S. A., Young, P. G. & Pivnicaworms, H. *Nature* **363**, 736–738 (1993).

factors downstream of Snf1p have already been defined by genetic screens (see, for instance, ref. 7). Another aspect of the phenotype of *snf1*-disrupted strains is that they fail to arrest at START in G1 phase when they run out of glucose⁸. This implies that Snf1p (and conceivably AMPK) inhibits progress through the cell cycle, as well as biosynthesis, in response to stress. Another possible connection is that the protein kinase most closely related to Snf1p in the fission yeast *Schizosacchar-*

omyces pombe is Cdr1/Nim1, although the carboxy-terminal domains are unrelated. Cdr1/Nim1 regulates the cell cycle in response to nitrogen starvation, apparently by phosphorylating the Wee-1 protein kinase^{9,10}.

On the other hand, the signal that switches the system on, and the mechanism by which the signal acts, is well understood in the mammalian system but unknown in yeast. Although changes in adenine nucleotide levels do occur when

yeast is starved for glucose, AMP does not appear to activate Snf1p. Nevertheless, it could be that some other small molecule acts as a starvation signal in yeast through a similar mechanism, and phosphorylation of the synthetic peptide now provides a biochemical assay to identify this putative signal molecule. □

D. Grahame Hardie is in the Department of Biochemistry, University of Dundee, Dundee DD1 4HN, UK.

OBITUARY

Julian Schwinger (1918–94)

I FIRST met Julian Schwinger in 1943 at the Massachusetts Institute of Technology. After bouncing back and forth a few times between Cambridge and Chicago, Schwinger had chosen — for his own reasons — to work on radar at the wartime MIT Radiation Laboratory and not, as he had been urged to do, on atomic weapon development in the Manhattan Project. Barely 25, he was already a legendary figure; I, two years younger, was an MIT graduate student about to submit my dissertation. Julian was half-a-dozen years away from his monumental work on quantum electrodynamics (the interaction between electrons and photons) which would bring him the Nobel Prize in Physics in 1965, a prize he shared with Feynman and Tomonaga.

Although Schwinger was world-famous and well travelled, he was quite unworldly. Bold and sure-handed in his physics, he was reserved to the point of shyness in ordinary daily life. He taught me much about physics; I like to think I taught him a little about getting along in the world. We became lifelong friends. Fifty-one years later he developed pancreatic cancer and his life precipitously ended, at the age of 76, on 16 July 1994. He continued working to the end.

Born in New York City, with absolutely no science in his background, he became committed to physics at an unusually early age. He once remarked that he had been reading straight through the *Encyclopaedia Britannica* and when he came to the letter P and to physics, that was it. Largely self-educated, he published his first paper at the age of 17, while an undergraduate student at City College of New York which he had entered the year before. His precocity caught the attention of I. I. Rabi (a future Nobel laureate), who brought him to Columbia University where he published his second paper, also minor, later that year. His first solo publication, on the quantum theory of the scattering of neutrons by magnetic materials, was another matter entirely. Submitted to the *Physical Review* on 11 January 1937, when he was still 18, it was an important piece of work, mature and elegant. The Schwinger techniques, his characteristic mastery of the subject and his physical insight, were on full display for the first time. He was, it was clear, a prodigy.

His progress in those early years was rapid and his contributions major, first to nuclear physics at Columbia and Berkeley, and then to radar at MIT.

After the war, Schwinger joined the faculty at Harvard and became a full professor at 29. Between 1948 and 1950 came the contributions that led to his Nobel Prize. His

beyond his published work. His lectures were elegant, lucid and original (he never did anything the same way twice), works of art and physics both. During the war years, I was asked to write up his lectures on radar waveguides, work we published as a monograph after the war, and so was a direct and happy beneficiary of his extraordinary qualities as a lecturer. Over the years, many others have similarly benefited from his lectures on such subjects as nuclear physics, quantum mechanics and field theory.

Even more important was his role as a mentor. He directed more than 70 doctoral theses, and his students (and their students) fill the faculties of America's best universities. Of Julian's students, three have also won the Nobel Prize: B. Mottelson and S. Glashow in physics, and W. Gilbert in biology. Schwinger gave his students much more than guidance on their research. He gave them a depth of understanding and a mastery of the field which permitted each to become not a Schwinger disciple, but an independent scientist. Through his students, he has had a more widespread and profound influence on theoretical physics over the past forty years than any other physicist.

Julian was a gentle and cultivated man. He and I had in common a serious love of music. I learned, mainly because of him, to play the recorder; he took piano lessons for decades. "If something is worth doing, it is worth doing badly", he once said of his playing.

There was a bitter-sweet quality to the later part of his life, for theoretical physics changed directions. It became, in Schwinger's view, too speculative, inadequately linked to experiment. He, accustomed to leading, chose not to follow. He became increasingly isolated, and, to a degree, estranged from the world community of physicists. Those of us who knew and loved him were saddened by this turn of events. But we took comfort in the certain knowledge that he was one of that handful whose magnificent contributions made science the great intellectual adventure of the twentieth century.

David S. Saxon

David S. Saxon is Professor Emeritus at the University of California Los Angeles, Department of Physics, Los Angeles, California 90024, USA.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Schwinger — Nobel prizewinner with Feynman and Tomonaga.

Lorentz and gauge-invariant reformulation of quantum field theory provided a new and self-consistent basis for renormalizing the formally ill-defined mass and charge of the electron. Contrary to what many believed, no new hypotheses were required, but only the ability to see and work through the extraordinary complexities and apparent contradictions inherent in the conjunction of special relativity and quantum mechanics. His work and Feynman's, over the next decade, revolutionized the quantum theory of fields and of elementary particles and laid the foundation for much of the subsequent spectacular progress in high-energy physics and the ultimate structure of matter. In 1972 he left Harvard to come to the University of California at Los Angeles, where he had the title of University Professor.

Julian Schwinger's legacy goes far

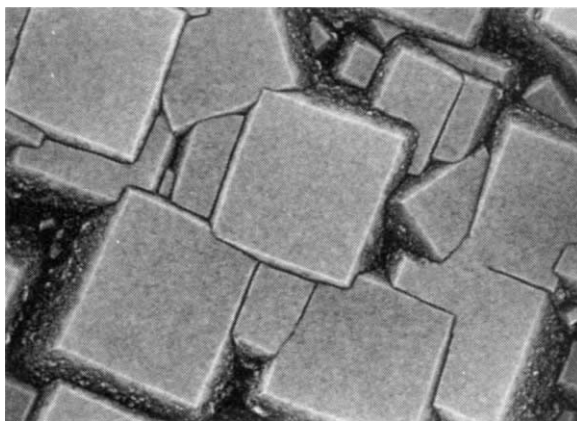
Electronic era elusive

John C. Angus and Alan T. Collins

DIAMOND seems to attract superlatives: the atomic density of diamond is the highest of any terrestrial material, and this, coupled with its strong, directional covalent bonding, gives it the highest known elastic modulus (hence its hardness), the highest room temperature thermal conductivity and the highest speed of sound. Less well known has been diamond's potential as a semiconductor. Diamond is structurally and electronically analogous to silicon. It is a wide-band-gap semiconductor and has an unusually high breakdown voltage and low dielectric constant. These properties, coupled with recent advances in the growth of diamond by chemical vapour deposition at low pressure, have led to speculation that diamond might find widespread application in high-speed electronic devices and devices designed to be operated at high temperature. Synthetic diamonds matching the quality and size of their natural counterparts have been a goal of materials scientists for years, and at the biennial diamond conference last month*, it seemed that this ambition has in part been achieved.

Predictions of a brilliant electronic future for diamond have, up to now, been unfulfilled because of the poor quality of synthetic diamond grown from the vapour and because of inherent problems with doping diamond to achieve appropriate conductivities. But researchers at Kobe Steel USA have been growing boron-doped, p-type synthetic diamond films on natural diamond substrates by chemical vapour deposition, and have produced films that have hole mobilities approaching those of the best observed natural diamonds. The mobility of charge carriers is a particularly sensitive indicator of the crystalline perfection of the host lattice (and is essential for effective conduction of charge). Their measurements showed hole mobilities of $1,400 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ at room temperature (J. Glass, Kobe Steel) which can be compared with typical mobilities in natural type II-b semiconducting diamond that range from 500 to $1,000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ (the best ever reported is $2,000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$). In addition, the level of undesired impurities that compensate the charge has been reduced to $5 \times 10^{15} \text{ cm}^{-3}$, equal to that in typical type II-b natural diamond, the standard for diamond quality.

One long-term objective of research on the electronic properties of diamond is the development of semiconductor devices that can operate at high temperatures, taking advantage of diamond's high band gap and thermal stability. The Kobe research group can now operate diamond field-effect transistors up to 400°C , and construct simple logic devices.



Oriented diamond crystallites grown by microwave-plasma-assisted chemical vapour deposition on silicon (100) after bias-enhanced nucleation. Further growth under conditions favouring (100) texture leads to highly oriented mosaic film. The crystal faces are about $1.5 \mu\text{m}$ square. (Deposition performed by Yaxin Wang, Case Western Reserve Univ.) Similar experiments on silicon carbide rather than silicon as a substrate produce still more highly aligned crystallites.

Diamond films of very high quality can also be grown on silicon carbide (Hiroshi Kawai, Waseda Univ., Tokyo). Although not technically single crystals, these films exhibit a mosaic structure in which the mean angular deviation of the individual crystals is only 2° . The key to these highly oriented films is a rather mysterious pretreatment of the substrate surface with an *in situ* bombardment obtained by imposing a negative bias of several hundred volts on the surface. This process, first introduced by Shigemi Yugo (Univ. Electro-Communications, Tokyo), has never been satisfactorily explained, but it provides a large number of diamond nuclei which are subsequently subjected to conditions that favour the growth of crystallites with their principal cubic axis perpendicular to the substrate. The result is a mosaic film approaching single-crystal quality. The figure shows an intermediate stage in the development of an oriented diamond film on the cubic face of silicon. The oriented crystallites have just started to grow together.

These achievements are particularly noteworthy because diamond synthesis by chemical vapour deposition is immature compared with silicon and gallium arsenide technologies. In addition, under

the growth conditions used for thin-film diamond, carbon prefers to exist as the thermodynamically stable phase, graphite. This makes it all the more difficult to develop synthesis methods and ion implantation processes for doping.

Not all of the reports were so positive. The goal of n-type diamond, important for device development, appears as elusive as ever. In n-type semiconductors current is carried by electrons donated by dissolved impurities that contribute a mobile electron to the crystal. From simple semiconductor theory the group V elements are

expected to form these donor centres in diamond. Nitrogen, which is adjacent to carbon in the periodic table, is readily incorporated into diamond, but the extra electron is too strongly bound to produce significant conductivity. Phosphorus, the next group V element, which is widely used for doping silicon, is too insoluble in diamond to act as an effective dopant. In principle, the solubility limit can be overcome by using ion implantation, but despite sophisticated ion implantation schemes (Johan Prins, Univ. Witwatersrand, South Africa; Rafi Kalish, Technion, Haifa, Israel), stable reproducible n-type doping has not been achieved.

There has been a history of over-optimistic predictions for the application of diamond as an electronic material and, despite the real progress reported at the Kobe meeting, there are no active diamond-based electronic devices being marketed (although there is a market for diamond substrates for heat removal from electronic components). Large-scale application of active components will probably require further improvements in quality and decreases in cost.

Future electronic applications may lie in other directions. The use of diamond as an electron emitter was the most promising application discussed at the Kobe meeting. Properly treated diamond surfaces show a negative electron affinity: mobile electrons in the conduction band are at higher energies than electrons at rest in the adjacent vacuum, so they can easily escape from the surface. Possible devices include flat-panel computer and television displays, discharge lamps, cathodes and electron multipliers. If suitable reproducibility, long-term stability and current densities can be achieved, the market for them could be enormous. □

John C. Angus is in the Chemical Engineering Department, Case Western Reserve University, Cleveland, Ohio 44106, USA. Alan T. Collins is in the Wheatstone Physics Laboratory, King's College London, Strand, London WC2R 2LS, UK.

*Fourth International Conference on the New Diamond Science and Technology, Kobe, Japan, 18–22 July 1994.

The sturgeons' plight

William E. Bemis and Eric K. Findeis

STURGEONS and paddlefishes — 27 Recent species in two families, Order Acipenseriformes — have an especial claim on the attention of biologists and the public. For biologists, acipenseriforms are a group that has been extant for a quarter of a billion years; for the public, they are prize exhibits in aquaria and, of course, are the source of caviar.

But the message last month from an international gathering* of researchers was unhappily familiar: wherever they occur in the wild, in North America, Europe and Asia, these fishes are in

Russia's continuing annual release of 100 million sturgeon larvae into the Volga cannot keep pace. In the Amur River, shared by Russia and China, kaluga (*H. dauricus*) were largely unmolested until a warming of relations brought intense commercial fishing.

In North America, Atlantic sturgeon (*Acipenser oxyrinchus*) never recovered from over-fishing a century ago, while others such as pallid sturgeon (*Scaphirhynchus albus*) were hit by river rechanneling. In Western Europe, the Baltic sturgeon (*A. sturio*) is nearly extinct. And in China, the Gezhouba dam on the Yangtze River virtually eradicated giant Chinese paddlefish (*Psephurus gladius*) (Wei, Q., Yangtze River Fisheries Research Inst., Shashi).

One problem is the biology of sturgeon: long-lived and slow to mature, they depend on large rivers, which are under environmental pressure worldwide. Their size and predictable spawning runs make them easy to catch even with simple gear. Only in remote areas (northern Ontario; M. Ferguson, Univ. Guelph), small, well-managed fisheries or under strict protection (*A. brevirostrum*; J. Boreman, Univ. Massachusetts, Amherst) have sturgeon avoided decimation.

Although we know something of the generic relationships within the acipenseriforms, the group's vast range and great variation complicate analyses at the α taxonomic level. The resulting disputes are not merely academic. For instance those centring on the Alabama sturgeon, *S. suttikusi*², generated editorials in the *Wall Street Journal*³. Is this a distinct species? Is it extinct? If not, should development of the Alabama River proceed given the sturgeon's warranted protection under the Endangered Species Act?

There are however rays of light in this gloomy picture. We finally have some basic life-history data on sturgeon, thanks to telemetry (B. Kynard, Univ. Massachusetts, Amherst), and international awareness of their plight is growing. Moreover, gourmets can do their bit by insisting on caviar from fish farms. Elimination of the wild harvest of this delicacy would help to ensure the survival of these ancient and remarkable fishes. □

William E. Bemis and Eric K. Findeis are in the Graduate Program in Organismic and Evolutionary Biology, University of Massachusetts, Amherst, Massachusetts, 01003, USA.

Diminishing resource: a large kaluga sturgeon from the Amur River, near Khabarovsk, Siberia. Kaluga are endemic to the Amur, but are fast becoming depleted in numbers not least because they yield high-quality caviar. (Photograph by V. Svirsky.)

trouble. At one time so numerous in North America that they posed hazards to navigation, all species now face extinction¹ because of over-fishing, poaching, pollution, habitat destruction and dams barring access to spawning sites.

Habitat destruction in the Aral Sea basin threatens the only surviving species of *Pseudoscaphirhynchus* (V. Birstein & B. Goncharov, Inst. Developmental Biology, Moscow). Uncontrolled exploitation of beluga (*Huso huso*) in the Caspian Sea also endangers this, the largest species, for

Deep heat

If you let a water drop fall on a red-hot plate, it does not touch the plate. Instead, it flies vigorously around on a cushion of steam. The steam seems to be generated preferentially towards the rear of the moving droplet, and pushes it along.

Daedalus once invented a red-hot boat based on this principle. Its hull, separated from the water by a film of steam, experienced almost no viscous drag. As it moved through the water the generated steam exerted an asymmetric thrust which helped it along, ramjet fashion. The principle might apply even better to a red-hot submarine. With no wave losses and no drag it might reach enormous speeds. Furthermore, below 2,200 metres the hydrostatic pressure exceeds the critical pressure of water. The vessel would be surrounded not by steam but by supercritical fluid of even lower viscosity. This would merge into the surrounding ocean, without the instability and rapid heat transfer of a discrete steam-water boundary.

At first, Daedalus feared that only a nuclear source could keep his submarine hot enough. But he then recalled that the ocean is in equilibrium with the atmosphere, and must be fully saturated with air. To be saturated at the local high pressure, the deeps must contain a vast amount of dissolved oxygen. Oxygen-saturated supercritical water, of course, supports combustion strongly. It is even used to burn up oily organic waste.

So Daedalus's red-hot submarine will take the form of an unmanned oil tanker. Its shape will be optimized for ram-steam propulsion; it will have a porous surface sealed with nitrocellulose lacquer. It will be towed out into the ocean by a tug, and allowed to sink. At the correct depth the lacquer will be fired, igniting the oil as it exudes from the uncovered pores. The resulting sheath of superheated steam and combustion products will stream back along the shaped hull, building up its thrust as a high-pressure external combustion supercritical ramjet. Eddies will not be shed from the smooth, compressible supercritical boundary layer. The only viscous drag will come from a small navigation package towed safely on a long cable behind the red-hot craft, which may reach speeds of hundreds of metres per second. Near its destination, the navigation package will steer it up towards the surface. Its propulsion will falter; the package will deploy floats and emit a homing signal for the collecting tug. Deep sea oil transport should be very clean. Leaks or even major spillages will never reach the surface, but will burn up in the oxygenated depths. David Jones

*International Conference on Sturgeon Biodiversity and Conservation, American Museum of Natural History/Hudson River Foundation, New York, 28–30 July 1994.

1. Birstein, V. J. *Conserv. Biol.* 7, 773–787 (1993).
2. Williams, J. E. & Clemmer, G. H. *Bull. Alabama Mus. nat. Hist.* 10, 17–31 (1991).
3. *Wall Street J.* A14, 22 Nov. 1993; A14, 26 Jan. 1994.

Carbon shells in flames

SIR — Fullerenes^{1,2} can be formed in hydrocarbon flames³. We have now found that carbon nanotubes^{4,5} and nanoparticles^{4,6-8}, along with soot particles containing fullerene-like shells, can also be formed in this way. Previously, such structures have been produced only by more energetic processes such as arc vaporization of carbon⁴⁻⁸.

Flames of acetylene, benzene or ethylene premixed with oxygen and a diluent were stabilized on a burner³ in a low-pressure chamber. Samples of condensable material were collected from within the flames using a water-cooled suction probe and from the water-cooled surface of the chamber under different sets of conditions over the ranges: burner chamber pressure, 20–97 torr; C/O atomic ratio, 1.06 (C₂H₂), 0.86–1.00 (C₆H₆) or 1.07 (C₂H₄); gas velocity at the burner (298 K), 25–50 cm s⁻¹; diluent concentration, 0–44 mol%; peak temperature 1,930–2,050 K.

The samples were analysed in a high-resolution transmission electron microscope (HRTEM) operated at 200 keV with a point resolution of 0.17 nm, a spherical aberration coefficient of 0.4 mm, and exposure times below the mini-

mum for electron beam damage. Each sample was dispersed in toluene using mild sonication. Drops of the dispersion were placed on holey carbon films on electron microscope grids and dried under vacuum.

Both probe and surface samples from the low-pressure acetylene and benzene flames were found to contain fullerene carbon structures consisting of nested shells separated by 0.34–0.36 nm, close to the interlayer spacing of graphite, with spheroidal, elongated and other shapes (*a* in figure). The nanostructures are mostly in the size range 2–15 nm, considerably smaller than the soot, which consists of particles of about 35–50 nm diameter in 100–3,000-nm agglomerates. The yield of nested nanostructures is typically ~10% of the soot yield and ~1% of the carbon fed in the benzene flames, somewhat less in the acetylene flames, and nil in the ethylene flames.

Some of the elongated structures are coaxial cylinders with spheroidal caps (*b* in figure), similar to multi-shelled nanotubes produced in arc-discharge graphite vaporization systems^{4,5}. However, the present results include fullerene tubes with both ends capped, reflecting growth

with both ends free. The nanotube in panel *b* of the figure appears to have grown from left to right, with more than one shell forming simultaneously.

Some of the nanoparticles are polyhedral, with smoothly curved continuous junctions between nearly flat faces that reveal fullerene structure. Others, also fullerene, are strikingly spherical, similar to some nanotube caps^{4,5} and nanoparticles⁷ produced by vaporization^{4,5} and electron-beam irradiation⁷ of carbon. Prolate spheroids are shown in *c* in the figure, the left one expanding upwards to the left with an angle of about 11°, consistent with the lower cap of each shell having five pentagons and the upper cap seven, implying an expansion angle of 20° or more⁵ orthogonal to the angle seen here.

Nanostructures observed but not shown here include nested conical tube tips similar to those seen in arc-discharge samples⁵, and curved layers or incompletely closed shells (fullerene carbon) that comprise the internal structure of the soot particles. The possible existence of shell structures in flame soot was suggested early in fullerene research⁹ but was not previously assessed by high-resolution electron microscope analysis of soot from fullerene-forming flames.

Considering the single-shell fullerene³ and curved polycyclic¹⁰ molecules previously quantified in these flames and the nested nanostructures and shells within soot particles observed here, fullerene shells are far more prevalent than graphitic flat sheets in the material of these flames. Thus carbon with curved layer structure can be formed not only under highly energetic all-carbon conditions of arc-discharges^{4-6,8} and electron-beam irradiation⁷, but also under the thermally milder and oxygen- and hydrogen-containing conditions that are found in flames.

Jack B. Howard

Department of Chemical Engineering,

K. Das Chowdhury

John B. Vander Sande

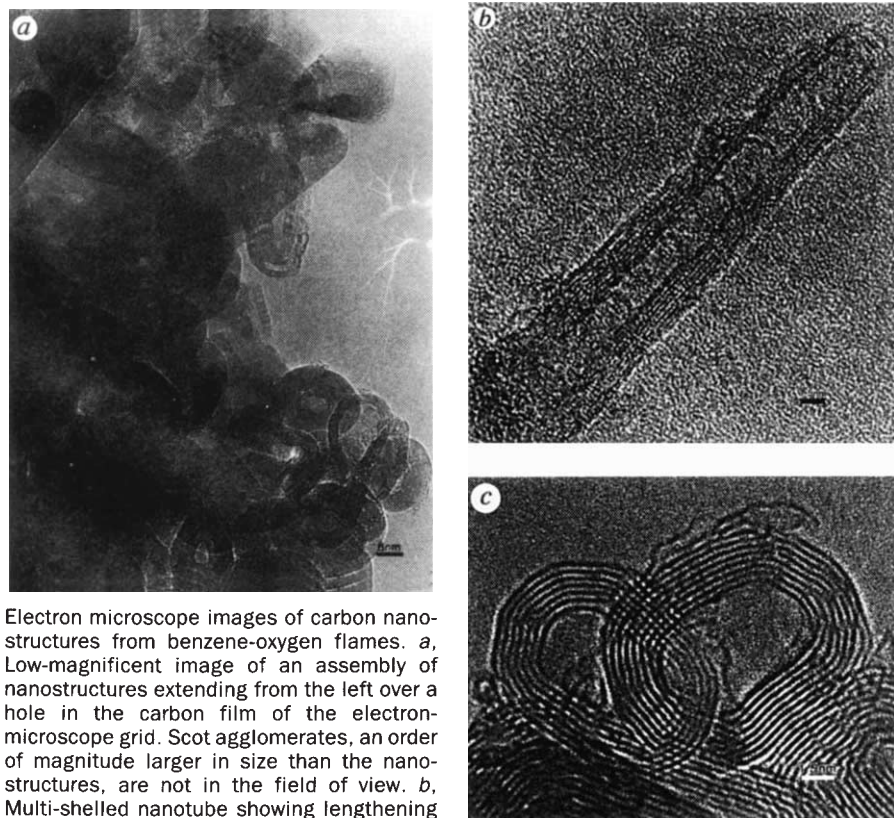
Department of Materials Science and Engineering,

Massachusetts Institute of Technology,

Cambridge,

Massachusetts 02139-4307, USA

1. Kroto, H. W., Heath, J. R., O'Brien, S. C., Curt, R. F. & Smalley, R. E. *Nature* **318**, 162–164 (1985).
2. Krätschmer, W., Lamb, L. D., Fostiropoulos, K. & Huffman, D. R. *Nature* **347**, 354–358 (1990).
3. Howard, J. B., McKinnon, J. T., McKinnon, J. T., Makarovskiy, Y., Lafleur, A. L. & Johnson, M. E. *Nature* **352**, 139–141 (1991).
4. Lijima, S. *Nature* **354**, 56–58 (1991).
5. Lijima, S., Ichihashi, T. & Ando, Y. *Nature* **356**, 776–778 (1992).
6. Lijima, S. *Cryst. Growth* **50**, 675–683 (1980).
7. Ugarte, D. *Nature* **359**, 707–709 (1992).
8. Ruoff, R. S., Lorents, D. C., Chan, B., Malhotra, R. & Subramoney, S. *Science* **259**, 346–348 (1993).
9. Zhang, Q. *et al.* *J. phys. Chem.* **90**, 525–528 (1986).
10. Lafleur, A. L., Howard, J. B., Marr, J. A. & Yadav, T. J. *phys. Chem.* **97**, 13539–13543 (1993).



Electron microscope images of carbon nanostructures from benzene-oxygen flames. *a*, Low-magnification image of an assembly of nanostructures extending from the left over a hole in the carbon film of the electron-microscope grid. Soot agglomerates, an order of magnitude larger in size than the nanostructures, are not in the field of view. *b*, Multi-shelled nanotube showing lengthening at cap and thickening radially. *c*, Multi-shelled nanoparticles consisting of nested prolate spheroids, tapered (left) and of approximately uniform cross-section (right). Experimental conditions: samples from burner chamber surface deposits: *a* and *c*, pressure $P = 97$ torr, C/O = 0.99 atomic ratio, velocity at burner (298 K) = 36 cm s⁻¹, 43.5% diluent (He/Ar = 2.6); *b*, composite sample from ranges of conditions given in the text. Scale bars: *a*, 6 nm; *b*, 1.5 nm; *c*, 1.5 nm.

Yucca sex

SIR — As Peter Moore reported in *News and Views*¹, there is indeed a “fly in the ointment” of the relationship between *Yucca* and its pollinator, and it is definitely a Dipteran. The close interdependence between the *Yucca* and its moth may involve a “balance of self-interests” as Moore notes, but there is plenty of opportunity for unbalanced positions. Large populations of moths may tilt the scales to one side and reduce or eliminate *Yucca* seed production. The unpredictable timing of viable fruit development may reduce such effects of moth overabundance². The scales will assuredly tilt the other way if moths are few, *Yucca* flowers many and pollination service is inadequate.

Such events are common in those portions of the *Yucca* geographical range where late and unpredictably cold springs decrease moth populations to low levels. Under these conditions, the fly *Pseudocalliope* sp. nov. (Lauhaniidae, Diptera) can be an important pollinator³. These flies exist in large numbers within flowers, where they court and mate. They carry *Yucca* pollen on their bristly bodies, from anthers to stigma both within and among flowers. The *Yucca* populations where this fly is found are also self-compatible, and can be pollinated via pollen deposition on the stigma (rather than requiring elaborate pollen-tamping behaviour by the moths into a stigmatic crevice as is often described⁴). In these populations, hand pollination can also increase seed set, in contrast to the populations described in ref. 2. The high densities and abundant reproduction of *Yucca* populations in these regions attest to the long-term effectiveness of this alternative solution.

These results indicate that *Yucca* “self-interests” are well served by keeping a flexible breeding system receptive to variable ecological settings, rather than depending on the conventions about the perfection of mutualisms.

Yan B. Linhart

Rhea J. Dodd

Department of Environmental,
Population and Organismal Biology,
University of Colorado,
Campus Box 334,
Boulder, Colorado 80309, USA

SIR — Moore in a *News and Views* article¹ compared how yuccas and figs control seed consumption by their pollinators. We would like to point out that several recent studies show that there is little or no empirical support for many of the mechanisms that have commonly been proposed to allow figs to avoid over-exploitation by their pollinator/seeds predators.

First, in many cases, the foundress wasps which pollinate the figs carry more than enough eggs to saturate all of the female flowers⁵. Second, style length within figs is distributed unimodally and not into short- and long-styled classes and the ovipositors of wasps are long enough to oviposit in a much larger proportion of flowers than actually develop into wasps⁶. Third, in many fig species the ostioles (the pore through which fig wasps gain entrance to the flowers) are simple structures that do not filter non-adapted pollinator species, nor do they prevent certain non-pollinating wasp species from entering the figs, laying their eggs, and then leaving⁷. Finally, most fig trees do not produce crops of fruit simply to maintain the wasps, the only documented example being *Ficus carica* (the edible fig)⁸. Overall, researchers' perceptions of the fig-wasp mutualism changed drastically as a result of studies done in the mid-1980s.

We hope that this shift, and its bearing on the more general understanding of mutualism, will soon be more widely appreciated.

F. Kjellberg

M.-C. Anstett

CNRS-CEFE, BP 5051,
34033 Montpellier,
France

E. A. Herre

Smithsonian Tropical
Research Institute,
PO Box 2072,
Balboa,
Panama

1. Moore, P. D. *Nature* **368**, 588–589 (1994).
2. James, C. D. et al. *Oikos* **69**, 207–216 (1994).
3. Dodd, R. J. & Linhart, Y. B. *Am. J. Bot.* **81**, 815–825 (1994).
4. James, C. D. et al. *Oecologia* **98**, 512–517 (1993).
5. Herre, E. A. *Experientia* **45**, 637–647 (1989).
6. Bronstein, J. L. *Ecology* **69**, 1298–1302 (1988).
7. Ware, A. B. & Compton, S. G. *Biotropica* **24**, 544–549 (1992).
8. Kjellberg, F., Gouyon, P. H., Ibrahim, M., Raymond, M. & Valdeyron, G. *Evolution* **41**, 693–704 (1987).

Painted out

SIR — Lister's evocative phrase in *News and Views*¹ that mammoths survived to co-exist with the Egyptian pharaohs led Rosen to suggest that an individual dwarf mammoth appears in a scene in a wall painting in a pharaonic tomb². Unfortunately, there are several objections to this intriguing view.

In the tomb concerned (that of *Rekhmire* in the XVIIIth Dynasty), the painted scenes were executed in around 1430 BC and, therefore, are younger than the lowest radiocarbon ages quoted for the dwarf mammoths of Wrangel Island in the Arctic Ocean³. The northeast Siberian specimens mentioned represented a dwarf form of the woolly mammoth, *Mammuthus primigenius* (Blumenbach);

although Rosen regards some elements in the creature illustrated as being suggestive of mammoth, certain diagnostic features (fatty hump, long pelt, small ears and distinctive tusks, curved in two planes) known for the type specimen of *M. primigenius*^{4,5} are entirely lacking.

Mention has been made that the bear in the painted scene would have been as exotic as a dwarf mammoth to the artist, but this applies equally to the elephant, for it was not native to Egypt in pharaonic times. Although the position of the tusks of the animal depicted is claimed as characteristic of a mammoth rather than of the African elephant, *Loxodonta africanus*, the upward-pointing tusks occur in stylized portraits of elephants whenever the latter appear in Egyptian art (and indeed in hieroglyphic writing, where the word for elephant, *abw*, includes an ideogram showing an elephant with upward-projecting tusks)⁶. Finally, although Rosen concedes that transport of live animals from central Africa to ancient Egypt has not been documented, he appears unworried that this also applies to the Arctic (after all, the bear accompanying the elephant in the tomb's tribute scene is not a polar bear).

Attractive though Rosen's theory may be, there is no reason to suppose that the Egyptian artist intended to depict anything other than an African elephant, the differential scaling of the human and animal figures being the result of stylistic convention rather than the naturalistic representation of extreme dwarfism. Similar intentional size disparities between human figures of different social rank are to be seen in Egyptian art and, indeed, within the same tomb.

William White

61 Eskdale Avenue,
Chesham,
Buckinghamshire HP5 3AY, UK

1. Lister, A. M. *Nature* **362**, 288–289 (1993).
2. Rosen, B. *Nature* **369**, 364 (1994).
3. Vartanyan, S. L., Garutt, V. E. & Sher, A. V. *Nature* **362**, 337–340 (1993).
4. Garutt, V. E., Gentry, A. & Lister, A. M. *Bull. zool. Nomencl.* **47**, 38–44 (1990).
5. Guthrie, R. D. *Frozen Fauna of the Mammoth Steppe: the Story of Blue Babe* (Univ. Chicago Press, 1990).
6. Gardiner, A. *Egyptian Grammar* 3rd edn (Oxford Univ. Press, UK, 1957).

Scientific Correspondence

Scientific Correspondence is a relatively informal section of *Nature* in which matters of general scientific interest, not necessarily those arising from papers appearing in *Nature*, are published. Because there is space to print only a small proportion of the letters received, priority is usually given according to general interest and topicality, to contributions of fewer than 500 words, and to contributions using simple language.

Purveyors of truth

Harry Collins

A Social History of Truth: Civility and Science in Seventeenth-Century England. By Steven Shapin. University of Chicago Press: 1994. Pp. 483. \$34.50, £23.95.

STEVEN Shapin is one of the foremost exponents of what has become known as 'social studies of science' or 'science studies'. The field blends history, philosophy and sociology of science into a critical understanding of how scientific knowledge is made. Shapin pioneered the combination of history and sociology; his example is increasingly influential. *A Social History of Truth* is Shapin at his best. It is a detailed and definitive historical study of Robert Boyle's establishment of experimental science, informed by an acute sociological sensibility. It is a scholarly *tour de force* that never loses sight of its theoretical purpose.

Shapin first reminds us that all societies rest on truth-telling and its correlate, trust. If people will as easily tell a lie as a truth, then nothing they say or do has meaning or consistency and there can be no social order. Samuel Johnson said that even the social order of Hell rests on trust. Scientific society is in no way special in this respect except in that it conceals its foundations. Second, Shapin argues that scientists must, perforce, work out ways of managing trust; they must work out whom to trust and whom not to trust. Third, in the bulk of the book, Shapin shows how Boyle manipulated the resources at his disposal — his gentlemanly standing — so he and other gentlemen would be seen as persons of a sort whose observations of natural phenomena should be believed.

Each element of the argument makes compelling reading. Shapin writes fluently, melding old and new sources and authorities seamlessly into the narrative. Like Robert Boyle, his control of detail is as masterful as his analyses are penetrating. Gentlemanly credibility rested on four qualities: gentlemen were taken to be competent sensory agents by virtue of their blood and their upbringing; gentlemen had to tell the truth to safeguard their reputations within a wide social circle; gentlemen were Christians; finally, gentlemen had the means to be free disinterested observers, unbound to any person or mundane end. The Royal Society's motto is *Nullius in verba* (take no one's word for it) but the word of gentlemen came to be accepted without strain. The mores of gentlemanly disagreement are

exposed and the parallels between the etiquette of 'giving the lie', which would lead to a duel, and challenging a claim about a matter of fact are drawn out. Shapin shows how philosophical violence was avoided by the use of what we would now call 'subhypotheses', which could

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Gentlemanly credibility — Robert Boyle (1627–1691).

save everyone's honour. Did a captain witness too buoyant an iceberg? Perhaps it was full of air, or the water was especially cold and dense!

Shapin shows that gentlemen might well pass the responsibility of experimentation and observation to assistants of lowly birth, but these people had no licence to speak the truth or to own knowledge; when their findings were uttered by gentlemen such as Boyle, however, the reports were credited with veracity. Technicians gained the 'invisible' status in the conduct of experiment that they have to this day.

In the epilogue, "The Way We Live Now", Shapin says that while today's sources of credibility may be different, the need for managing trust remains the same. The sheer scale of science suggests we must nowadays give our credibility to certified institutions rather than to individuals, but Shapin argues that science is still done in 'core-sets' where specialists know each other personally and make

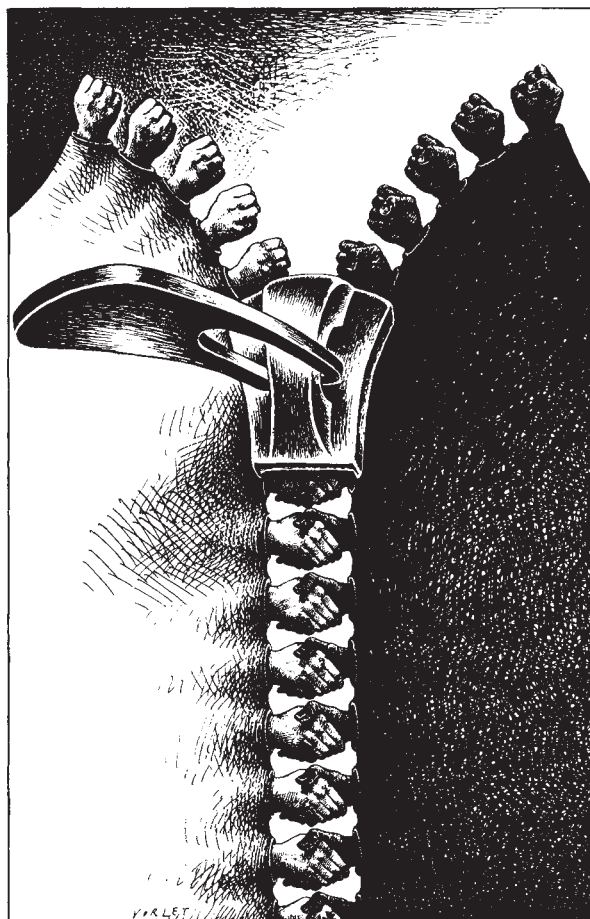
their judgements of credibility on the basis of private assessments of competence and on local reputation. "Scientists", he concludes, "know so much about the natural world by knowing so much about whom they can trust."

That Shapin's account of science is correct is hard to doubt. We know that despite science's most prevalent image of itself, experiments can be repeated only by those with the appropriate skill; to see a disputed result tested requires that one *knows* who has the right skill, and this cannot be measured; so witnessing an experiment may tell one much or little, depending on whether the demonstrator is an honest and competent scientist (and we can never be sure of even our own competence) or a dupe, dissembler or magician. In the light of Shapin's analysis we can see some of the recent fashion for 'vigilante-style' exposure of scientific heretics as a misunderstanding of the way the thread of trust holds science together (pulling too hard threatens to destroy the whole garment). We can see why the sometime-chemist and former British prime minister Margaret Thatcher, accepting the self-image of her old discipline, could think that self-interest and accountability were a substitute for truth, trust and society.

Shapin spends time defending the title of his book and it has to be said that it is a little 'cheeky'. That there cannot be social relations without truth and trust is a matter of 'acting truly'; it is a matter of constancy and reliability irrespective of what we believe. A history of truth would be about the historical tension between narrow self-interest and the development of

human societies. (It is worth noting that ants, bees and computers manage things differently.) There is no social history of truth, for a history of truth is a history of society. Trust may be everywhere necessary, but normal life and 'normal science' do not need trust management; trust management is Shapin's topic. In other words, trust management is to truth as good looks are to sex; good looks are not a condition of reproduction (see p. 193, where the ineradicable role of trust is taken to imply the necessity of an answer to the question 'whom to trust?'). Thanks to Shapin, we now understand how the managed plausibility of gentlemen in seventeenth-century society gave rise to our beliefs about what is in the world. Anyone who wants to understand the origins and meaning of modern science should read this book. □

Harry Collins is in the Science Studies Centre, University of Bath, Bath BA2 7AY, UK.



ONE world — this drawing by Christophe Vorlet, accompanying a commentary about the integration of traditionally black US colleges, appeared in *Chronicle of Higher Education*. It is taken from *Zipper: An Exploration in Novelty* by Robert Friedel (Norton, \$23). The author, a historian of technology and science, tells the story of the evolution of this gadget, which was patented in the late nineteenth century. He traces fashion fluctuations, design failures and marketing struggles as well as the colourful cast of characters behind them, while showing how the allure of the lowly clasp came to be reflected in films, songs and literature.

Angel in America

Richard Davenport-Hines

My Own Country: A Doctor's Story of a Town and Its People in the Age of AIDS. By Abraham Verghese. Simon and Schuster: 1994. Pp. 347. \$23*.

SINCE the identification of AIDS in the 1980s, there has been a plethora of general books by people touched by the syndrome. Many of these have been by patients or their survivors, but *My Own Country* is the first written by a physician about his patients in the Bible Belt of the United States. As a piece of medical literature intended for general readers, it bears comparison with the essays of Sir Frederick Treves or Sir William Osler.

In 1985, Abraham Verghese was a young doctor specializing in infectious disease and living in Johnson City, Tennessee. His book opens with a description of the first patient with AIDS in the city: a young man returning from New York to visit his parents, who dies of pneumocystis carinii, and whose illness causes such panic among the hospital authorities that, after his death, it was proposed to bury the

respirator to which he was attached in a swamp: as a compromise the machine was dismantled, and cannibalized for parts. AIDS was then seen as "a big city problem", but an increasing number of people diagnosed with HIV began presenting themselves to Verghese, who unexpectedly found himself the local AIDS expert.

He was initially a man nervous of male homosexuality and set on a safe, conventional medical career. He found himself deeply moved by his experiences, however, gradually taking detailed biographical notes not only from his patients but also from their families, friends and lovers. In some cases he discarded the conventional relationships of patient and physician, and became intimately involved. He witnessed acts of great kindness, courage and generosity from patients, carers, neighbours and a few nursing colleagues. He was also confronted by acts of exceptional stupidity, rudeness and cruelty (some regrettably from medical colleagues). "Life speeds up and heightens in climates of extreme fear and emotion", he concludes. "It is hard to live in these circumstances, despite the acts of tenderness that lighten everything." Much of this book is a memoir of individual patients, but he gives a transcendent significance to the stories he tells. He raises issues of medical practice, spiritual need,

social justice and sexual conduct, and sometimes gives his own answers, in a tone that is never peremptory or arrogant but has a gentle persistent subversiveness.

The AIDS patients whose cases he describes include gay men, intravenous drug users, haemophiliacs, the wife of a bisexual, a businessman infected by blood transfusion, and the wife whom the latter too has infected. He writes of them with a vividness that is an act of commemoration; he is never mawkish and keeps a measure of dignity even for patients far gone in dementia. His grief over the plight of some patients, his alertness to human nuances, his curiosity and capacity for surprising himself are all powerfully conveyed, although this is the least egocentric of books.

One of its strongest features is Verghese's gradually developing understanding of the diversities of human sexual behaviour. One woman nursing her brother says of him to Verghese, "Gay may have been what he did, but it wasn't who he was", and this remark might serve as an epigraph for his book. Any physician interested in the realities of Western sexual behaviour, rather than the sterile statistics of sexologists or the stereotyped polemics of militants, will find Verghese's ideas fascinating. He comes to suspect that gay men "are more representative of men than heterosexual men" because "the sexual activity of gay men, their sexual drive, the number and variety of partners, the ready possibility of anonymous sex, might represent what *all* men want, except that they can't get women to agree". Not surprisingly the strain that his new medical interest places on his marriage provides an ominous personal note.

My Own Country is outstanding. It is elegantly written, with descriptions of Tennessee landscape worthy of a Pulitzer prizewinner. It has grace, moral power and a strong social commentary. The lives of the Tennessee people seem dominated by technology — one dying man keeps the police radio monitor blaring all day in his bedroom — yet they remain as simple and even superstitious in their concerns as mediaeval serfs, although a patina of education conceals this from themselves and others. Verghese deals at times with highly emotive topics, and describes some deeply emotional scenes of life and death, but his literary treatment is pitched at just the right level of mild understatement. This is a modest, humane, reflective book, full of images and ideas that resonate and enrich. Every reader will be provoked to rethink old assumptions by this gently challenging and finely observed book. Every patient faced with serious illness will hope to have a physician like Verghese. □

Richard Davenport-Hines is at 51 Elsham Road, London W14 8HD, UK.

*Soundings in the UK (Phoenix House, £18.99).

Building a safer future

N. A. F. Smith

Design Paradigms: Case Histories of Error and Judgment in Engineering. By Henry Petroski. *Cambridge University Press*: 1994. Pp. 209. £30, \$42.95 (hbk); £12.95, \$17.95 (pbk).

HENRY Petroski's thesis is that the study of case histories is the key to understanding engineering failures and that these in turn shed light on the process of design, this being the very essence of engineering, the central feature that above all else imparts to the discipline a distinctive intel-

justify the study of engineering history not for its own sake but because it can monitor contemporary design or, better still, guarantee a safe future. Thus engineers, young and old, should study history for material, professional reasons.

The design problems examined by Petroski range from antiquity to modern times and each represents a paradigm of a type, such as the problem of a mistaken concept, limits to size, design change for the worse, logical errors and false success. Although it has been done before, the job is nicely done and the case for using engineering history to comprehend the nature of design is well made.

There are at least two aspects of Petroski's argument about which basic questions can be asked, however. Even when the nature of historical failures has been analysed fully and categorized by type, how do designers at the drawing board or in front of a computer screen actually recognize when, or indeed if, they are up against a design paradigm of one sort or another? In short, how is hindsight turned confidently and accurately into foresight?

Mary Evans

Second, leaving aside some Galilean examples that are of doubtful relevance anyway, the case histories discussed are drawn wholly from structural engineering and predominantly from bridge-building. Petroski acknowledges in the introduction that other engineering fields merit consideration, but until they are examined with comparable thoroughness, how far can we be convinced by a general philosophical position developed from only a few, specialized examples within one branch of a very large subject?

The author recommends his book for engineering students, practitioners and the general reader. The

general reader will be well served, especially because the case histories are so interesting and well presented in themselves, never mind the grander implications. Practitioners might reckon they are already thoroughly familiar with such philosophy of design. For students the suitability is more difficult to judge. Until one has experience of history and design, especially the latter, Petroski's ideas might not be readily appreciated. □

N. A. F. Smith is reader emeritus in the history of technology, Imperial College, London SW7 2AZ, UK.

Negative synergy

Charles Perrow

Learning from Disasters: A Management Approach. By Brian Toft and Simon Reynolds. *Butterworth-Heinemann*: 1994. Pp. 140. £25, \$49.95.

THE field of risk analysis and accident prevention has grown as fast as the disaster rate in the past 15 years. One branch is technical, involving fault-tree analysis and mathematical models of probabilities and consequences; organizations, if mentioned, are assumed to be highly rational. Another branch has been far more sociological and political; organizations achieve only limited rationality, and an analysis of power and group interests is required to account for the failure to use common sense.

This book takes an interesting middle position. Accidents are sociotechnical in nature, rarely the result of technical failures, but the analysis of the sociotechnical system is determinedly 'scientific' and rational, complete with extensive taxonomies and definitions. It draws on systems theory in the social sciences, popular in the 1960s. Elaborate boxes of system components arranged in a time flow pin down errors and point to remedial actions.

The results will be disappointing to anyone who now takes it for granted that 'acts of God' are not responsible for accidents, that systems are complex, that accidents have multiple origins and that preparedness and vigilance are essential. The reader will not find much more. The elaborate taxonomies make all accidents roughly similar — they are derived from public inquiries into 19 accidents in the United Kingdom between 1965 and 1978. Furthermore, the emphasis leans toward human failures rather than failures of organizational structure. The analysis is prosaic, the writing pedantic and the recommendations heroic — put safety first, find the money to pay for it, plan ahead, communicate with every conceivable organization, be eternally vigilant, and honest when caught.

Nevertheless, there is a reasonably interesting "schematic report analysis diagram" that analyses the Cambrian Colliery accident of 1965, based on the work of B. A. Turner (as is a great deal of the book), a pioneer in the accident field. It outlines the numerous failures and shows how the investigating committee ignored some of the more important ones. The diagram is useful for investigating committees, but when enlarged as a generic blueprint for an "organizational learning system", as it is in the final chapter, it mimics the failure of 1960s system theory: everything is (equally) important, connected and must be taken into account.

Collapse of a church in Brousse, France (1897).

lectual and creative character. In truth it is not a new proposition, especially for civil engineering; numerous earlier writers have followed a similar approach. In fact, something of the sort has, over the years, been implicit in all manner of historical and biographical writing simply because it is natural, even essential, to scrutinize such human activities as politics, warfare or industry through the lessons of individual or corporate misjudgement or failure. To this extent, Petroski's position is neither radical nor original.

His approach also seems to be, to a degree, contrived in that it is seeking to

Given the strides that have been made in understanding why organizations persistently refuse to 'learn' from mistakes, and the role of organizational interests and power in catastrophes, it is surprising that in their models, the authors assume that organizations will behave rationally, so that everything can be fixed in order to achieve high reliability. A long list of accident studies they do not cite challenges this assumption, most strikingly Scott Sagan's *The Limits of Safety* (Princeton University Press, 1993; for a review see *Nature* 367, 30; 1994), and before that the work of Lee Clarke, Diane Vaughan, Kai Erikson, J. S. Kroll-Smith, Paul Shrivastava and many others.

To be fair, there are references to the "bounded rationality" school of organizational analysis (James March) and to James Reason who is sceptical about rational models. But examination of the five accidents reported in detail reveals that it is unlikely that any of those responsible would have behaved differently had they carefully studied this book. Accidents are rare; other problems and interests are far more pressing than pursuing

the culmination of this book, the cumbersome "socio-technical failure minimizing systems model".

The puzzling thing about accidents and especially large catastrophes is that they are so few, given the enormous number of systems operating without the requisite structures so exhaustively enumerated here. It is hard to have just the right combination of system failures and environmental conditions to kill a lot of people in one blow. For this reason, countless risky systems successfully skirt disaster, skimping on safety and putting other interests first, while regulators fuss. The few that do experience "negative synergy", such as those described in this book, provide only obvious lessons that the rest continue to ignore for quite understandable if often deplorable reasons. Disturbingly, we have little more to say about the matter. □

Charles Perrow is in the Institution for Social and Policy Studies, Yale University, Box 1965, New Haven, Connecticut 06520-1965, USA, and author of Normal Accidents (BasicBooks, 1984).

The new geology

Andrew Goudie

Geology and Environment in Britain and Ireland. By Nigel Woodcock. UCL Press: 1994. Pp. 184. £14.95 (pbk).

THIS slim volume, aimed at first-year undergraduates and college students, attempts to outline the ways in which geology relates to human society, and to illustrate this with examples from the British Isles. It covers geological influences on society (landscape, human geography and hazards); Earth resources for society (land, water, construction materials, minerals, coal and peat, petroleum and energy); and human impact on Earth (resource extraction, engineering geology, and pollution).

This is an awful lot of material to compress into such a small compass. As a result, nothing is explored in detail and the style is compact to the point of being terse. On the other hand, the amount and style of illustrative material is first class, and I found myself spending far more time analysing the figures than in absorbing the text. The illustrations have all been specially drawn to a particular style and are refreshingly clear and simple yet informative. Students will also benefit from the guides to further reading, which are placed at the end of each chapter and include references to the geographical as well as the environmental literature.

The publisher makes certain claims for the book in terms of the changing nature of geology: "This book is published at a pivotal point in the history of geology. Scientists who, for a century and a half, have been preoccupied with finding earth resources are increasingly being asked where on earth to dispose of the effluents from using them . . . [the] final chapter breaks new ground in opening a debate on the ethical basis of applied geology — a debate which is needed to steer the subject into the 21st century."

Indeed, the author himself is somewhat damning about the level of environmental concern shown by geologists, which for the first half of this century is described as being "so thin that it can be recognized only in retrospect. Even the more recent growth in anxiety over human impacts on the environment has yet to filter deeply through the geological profession." The plea is that geologists should take a more holistic and systems-oriented approach to the environment. Following Woodcock's example they should read more geology. □

Andrew Goudie is in the School of Geography, University of Oxford, Oxford OX1 3TB, UK.



GIANT bellflower
(*Campanula latifolia*).
Oil painting by
Raymond Booth
(1969). The picture is
reproduced in *The Art
of Botanical Illustration*
by Wilfrid Blunt and
William T. Stearn
(Antique Collectors'
Club, £29.95). First
published in 1950, this
classic work has been
long out of print. Now
substantially revised
and enlarged, the new
edition contains 126
colour plates and 140
black-and-white
illustrations, and aims
to provide a
comprehensive and
critical survey,
spanning more than
3,000 years, of the
portrayal of plants.

The case for an open Universe

Peter Coles & George Ellis

Contrary to popular belief, there is as yet no convincing evidence that the Universe has a density near the critical value required for closure. Indeed, a low-density Universe is favoured by many of the observations. Cosmologists should therefore keep an open mind about this question until it is settled by empirical data.

THE issue of the average density of matter in the Universe has been a central question in cosmology since the first quantitative cosmological models were devised. In particular, cosmologists wish to know whether this density is above or below the critical density ρ_c , separating those models that recollapse in the future from those that expand for ever; in terms of the Hubble constant H_0 :

$$\rho_c = \frac{3H_0^2}{8\pi G} \approx 1.9 \times 10^{-29} h^2 \text{ g cm}^{-3} \quad (1)$$

where h is the Hubble parameter in dimensionless form and $H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1}$; the uncertainty in h is discussed below. The standard parametrization of the density^{1,2} is in terms of its ratio Ω_0 to this critical density:

$$\Omega_0 = \left(\frac{\rho}{\rho_c} \right)_{t=t_0} = \frac{8\pi G \rho_0}{3H_0^2} \quad (2)$$

where the subscript 0 indicates evaluation at the present time, that is, $t=t_0$. Then $\Omega_0 > 1$ if the Universe will at some future time collapse into a second singularity at a 'big crunch', whereas if $\Omega_0 < 1$ the Universe will expand for ever.

The critical-density case, which just succeeds in expanding for ever, corresponds to the Einstein-de Sitter Universe models with flat spatial sections, and occurs when $\Omega_0 = 1$. In recent years, largely as a result of the inflationary Universe models, there has been a large body of opinion claiming that the density is very close indeed to that in a critical-density Universe: $\Omega_0 = 1$ to very high accuracy today, despite the fact that until very recently all the unambiguous astronomical density determinations have given a much lower value. (The directly observed luminous material in the central regions of galaxies contributes only a very small fraction of the critical density: $\Omega_0 \approx 0.005$. Dynamical estimates of the masses of galaxies using, for example, rotation curves for spiral galaxies, suggest that galaxies contain large quantities of dark matter which might take the cosmological density up to $\Omega_0 \approx 0.04$ or so^{2,3}. Both these rough estimates are uncertain by a factor of at least h^{-1} , but are clearly less than unity. We shall discuss the evidence on Ω_0 from larger scales below.) This is the famous dark matter problem, the search for the so-far undetected matter that must be there if indeed we live in a critical-density Universe.

Although laboratory searches for dark matter candidates have so far been in vain, in the past few years this theoretical predilection for a high-density universe has been reinforced by various astronomical observations which have been taken as supporting the critical value.

In our view the case is far from closed. We believe it is an appropriate time to review the evidence, and to see if the case for a critical-density Universe is compelling. *Inter alia* the purpose of so doing is to emphasize that this is indeed an experimental question, where theory—no matter how dear it may be to us—will eventually have to bow to the experimental evidence. It may be that the theoretical prejudice in favour of the high-density models will one day be confirmed; if so that will be a great triumph for theory. However, at present the weight of evidence

if anything favours a low-density Universe. This conclusion supports the view of Peebles expressed some seven years ago³, despite the new observations that have accumulated in the intervening period.

We should mention at this point that there is a related issue, also controversial, namely the question of the possible existence of a cosmological constant, known as the Λ -term, in the cosmological field equations. This term corresponds to a vacuum energy density which affects the spatial curvature of the Universe without changing Ω_0 . In its presence, it is possible to have a flat Universe with $\Omega_0 < 1$ which expands forever; if $\Lambda = 0$ then $\Omega_0 < 1$ requires negative spatial curvature. We shall adopt an agnostic stance on this issue, and will not address the various arguments about Λ in any detail for reasons of space; we shall, however, raise the possibility in discussing those issues where it is related to the Ω_0 issue.

Theoretical arguments

The primary reasons for the widespread belief in a critical density of matter are theoretical. From the 1930s to the 1970s there was a tendency to prefer the Einstein-de Sitter (critical density) model simply because—consequent on its vanishing spatial curvature—it is the simplest expanding Universe model, with the simplest theoretical relationships applying in it. Thus it is the easiest to use in studying the nature of cosmological evolution. Moreover it is known that on the cosmological scale, spatial curvature is hard to detect (indeed we do not even know its sign)—so the real value must be close to zero. The pragmatic astrophysicist thus uses the simplest (critical density) model as the basis of his or her calculations—the results are good enough for many purposes. There are, however, other arguments for preferring Ω_0 close to unity on theoretical grounds. We shall discuss two of them here.

The copernican principle. The density parameter Ω evolves with cosmic time; at an arbitrary time $\Omega = 8\pi G \rho / 3H^2$, with H the instantaneous value of the Hubble parameter. In a model containing ordinary matter and/or radiation the cosmological expansion is decelerating, and the quantity $|1 - \Omega|$ increases with time unless $\Omega = 1$ exactly, in which case the equality holds for all time. This simple dynamical behaviour leads to a well-known conundrum. If $\Omega_0 \neq 1$, then Ω must have been closer and closer to unity the further we look into the past. How did such a fine-tuning arise? To put this another way, consider the radius of curvature of the Universe, which is determined by Ω_0 . If Ω_0 is not too far from unity at the present epoch then the present curvature radius is close to the Hubble radius, c/H_0 . There seems to be no *a priori* reason why we should live at an epoch when this coincidence holds if Ω is evolving with cosmic time.

This coincidence would seem to violate the most general form of the copernican principle if the Universe is open; that we do not occupy a special place in the Universe (the argument here applies to temporal, rather than spatial position). On the other hand, if the Universe is flat, the copernican principle is not violated because Ω does not evolve with time. This argument therefore seems to suggest one should prefer a flat universe, if one accepts the copernican principle.

There are two important objections to this line of argument. First, there is the problem of interpreting the 'coincidences' noted above. Arguments of this type are probability-based (coincidences are events which seem 'improbable', given what we think we know about the circumstances in which they arise). At the present time, however, we do not know how to define a unique measure on the space of allowed cosmological models⁴⁻⁷. We are not, therefore, in a position to make meaningful probabilistic arguments about this question, except that an exactly flat universe is presumably infinitely improbable. Second, one should always remember that we do to some extent occupy a special temporal position, even in a flat universe, if only because of the requirement that there has to be enough time between the Big Bang and the present epoch for life to evolve⁸. This observation is an example of the use of the weak anthropic principle⁸; though often dismissed by cosmologists, this kind of argument is actually essential to the correct formulation of probabilistic arguments in cosmology⁹. The only models in which the copernican principle, in the sense used here, is exactly true are the steady-state cosmologies.

Inflation. The inflationary universe scenario¹⁰⁻¹³ has provided a new and much stronger theoretical motivation for critical-density models. In inflation, unlike the usual models, the cosmological expansion is accelerated for a time, and in this period $|1 - \Omega|$ decreases; so if inflation lasts long enough, then Ω_0 must be arbitrarily close to unity. This argument is frequently advanced by cosmologists motivated by particle-physics considerations, but there are several reasons why it should be treated with care.

First, there are indeed inflationary models that lead to present-day values of Ω_0 other than one¹⁴⁻¹⁷. Their existence is shown either by running physics backwards from any chosen value of Ω_0 today, or considering^{7,18} the (Ω, S) phase planes for inflationary cosmologies (where $S(t)$ is the Robertson-Walker scale factor). Furthermore, there are classes of universe (power-law inflation, the coasting universe families) which are essentially inflationary but do not drive Ω to unity at the present epoch, as standard inflation does^{14,17}. One condition necessary for the construction of such models is that they do not obey the usual 'slow-rolling' condition (that the quantum field driving the inflationary epoch evolves slowly compared with the expansion rate of the Universe). In these models it is true that a value of, say, $\Omega_0 \approx 0.2$ (at the present time) does require some degree of 'fine-tuning' of the conditions before inflation, but it is by no means clear how strong an argument this is against them. The fine-tuning argument is inherently probabilistic, and should therefore be phrased in terms of some well-defined invariant measure of probability of the space of universe models; we do not at present have such a measure⁴⁻⁷. The 'obvious' measure often used gives results that are extremely time-dependent: for example, a low-density Universe today comes from values of Ω very close to zero at the beginning of inflation—but very close to one at the Planck timescale, $t = t_p \approx 10^{-43}$ s, which is usually thought to be the limit of applicability of arguments based on classical gravity. The problem is that this measure is not invariant as the universe evolves. A rigorous version of the inflationary argument based on well-defined notions of probability and measure is yet to be produced.

In any case, even the 'standard' inflationary models do not necessarily drive the value of Ω_0 to unity. What they do is drive the spatial curvature at the present epoch to zero. If one is prepared to accept the existence of a Λ -term then one can have flat spatial sections, and $\Omega_0 < 1$.

The possibility that an open universe might arise from inflation is controversial, and while we believe the arguments above should be taken seriously, we do not intend to hang our case upon them. In the end, our main point about inflation is that we do not in fact have any proof that inflation ever took place in the early Universe (nor do we even have a well-defined candidate for the 'inflation' scalar field). Even if one rejects the ideas

we discuss above, the argument from inflation should not itself lead one to conclude that Ω_0 is near unity, for it may never have happened.

Whatever one thinks about theoretical arguments such as those from inflation, our uncompromising position is that, in the end, theory must give way to observation. The history of astronomy and cosmology is replete with examples of strongly held belief eventually turning out to be wrong. An important example from the early days of modern cosmology is the theoretical conviction that the Universe must be static, unquestioningly held by all the major workers at the time. Theory must ultimately be tested by whether or not observational evidence confirms its predictions. We now turn to that evidence.

Cosmological observations

The basic properties of the Universe, such as its age, geometry and composition, depend on the parameter Ω_0 . These properties therefore provide a way to constrain Ω_0 relatively directly. In this section, we look at examples of this kind of argument and show that the evidence does indeed point to a low-density Universe.

The age of the Universe. An obvious way to test a cosmological model is to compare the predicted age of the Universe with the inferred ages of astronomical objects¹⁹⁻²³. One can date lunar and meteoritic rock using long-lived radioactive isotopes such as ²³⁵U. As these heavy isotopes were presumably formed in supernova explosions, modelling of the rate of such explosions allows one to estimate the age of the Galaxy. Typical numbers quoted for this type of analysis are $t \approx (9-16)$ Gyr, the range in estimates mainly representing uncertainty in the modelling²⁰. Stellar evolution also provides stringent limits on t_0 . The very oldest stars are observed in globular clusters and these objects can be dated fairly accurately by fitting isochrones to the Hertzsprung-Russell diagrams of the stars they contain. Ages have been quoted up to 20 Gyr for some of the oldest clusters, but many of these claims are still controversial. The most likely range of ages is $t \approx (14-16)$ Gyr. For globular clusters to be significantly younger than this there would have to be a big error in our understanding of stellar evolution.

We know that the Universe is expanding in such a way that the apparent recession velocity (v) of a distant galaxy is proportional to its distance (r): $v = H_0 r$. The characteristic time for the expansion of the Universe is therefore defined by the Hubble time, $\tau = 1/H_0$. In terms of h , the Hubble time is roughly $10h^{-1}$ Gyr. It is believed that h lies in the range $0.4 < h < 1$. However, the Hubble time τ is an overestimate of the actual age of the Universe, t_0 , which depends on Ω_0 also. For an Einstein-de Sitter universe, $t_0 = 2\tau/3$ whereas for $\Omega_0 = 0.1$, $t_0 \approx 0.9\tau$.

Even the most conservative interpretation of the estimates of astronomical ages would lead one to conclude that $t_0 > 13$ Gyr. This is already in conflict with the Einstein-de Sitter universe unless $h < 0.5$. If it turns out that $h \approx 0.8$, then we will have to abandon the idea that $\Omega_0 = 1$. For example, a universe with $\Omega_0 = 0.1$ and $h = 0.5$ has an age of 18 Gyr, comfortably in accord with age estimates. If $h = 0.8$ this reduces to 11 Gyr, so that a high h is also a problem unless Ω_0 is very small indeed. (Even an empty universe would be only 16 Gyr old if $h = 0.8$, so if the age estimates turn out to be conservative, and $h \approx 0.8$ there seems to be no alternative but to introduce a Λ -term, which can increase t_0 relative to τ). The critical-density universe is therefore in great difficulty if h is towards the upper end of the usually-accepted range. The evidence with regard to h has recently been reviewed² and we believe that, although the situation is far from clear, the weight of evidence is indeed beginning to shift towards $h \approx 0.8$ rather than 0.5.

Element abundances. One of the great successes of the hot Big Bang model is the agreement between the observed abundances of light elements and calculations of nucleosynthesis in the primordial fireball. Current estimates of the relative primordial abundances of deuterium (²H), ³He, ⁴He and ⁷Li all agree with

nucleosynthesis predictions²⁵, only if the density parameter in baryonic material is in the range

$$0.01 < \Omega_b h^2 < 0.015 \quad (3)$$

A baryon density higher than this would produce too much ^4He and ^7Li , while a lower value would produce too much deuterium and ^3He .

That nucleosynthesis favours a low value of the baryon density is not a new conclusion²⁶, but many of the parameters needed for these calculations, such as the neutron half-life and the number of light neutrino types, are much better known today from direct particle experiments. The discrepancy between the value (3) and $\Omega_b = 1$ is the main argument for the existence of non-baryonic dark matter if the Universe has a critical density. Without postulating the existence of such material, one is led to a low value of Ω_0 . If one admits the possibility of $\Omega \approx 0.99$ of non-baryonic cosmological matter to resolve this, there are problems in accounting for some astrophysical phenomena. In particular, a recent analysis of the Coma cluster²⁷ indicates that there are simply too many baryons there compared with the amount of dark matter, to square the constraint (3) with $\Omega_0 = 1$ globally. A strong limit on Ω_0 emerges:

$$\Omega_0 \leq \frac{0.15h^{-1/2}}{1 + 0.55h^{3/2}} \quad (4)$$

Note that even if $0.1 < \Omega_0 < 0.3$, one still requires some of this to be non-baryonic to be compatible with standard nucleosynthesis and the favoured values for h . Various attempts have been made to produce models of nucleosynthesis that can accommodate a higher value of Ω_b , notably those in which inhomogeneities arise during the quark-hadron phase transition²⁸. None of these seems to be able to reconcile a critical-density baryon-only universe with the observations. It is less clear whether it may be possible to accommodate $\Omega_b \approx 0.2$ by these means.

'Classical cosmological tests'. In the early days of observational cosmology, much emphasis was placed on the geometrical properties of expanding universe models as tools for estimating the density parameter^{2,29}. It was later realized, however, that there is strong evolution of the properties of astronomical objects that might be used to trace the geometry, and that this evolution usually dominates over geometrical effects. A brave effort has been made recently by Kellerman³⁰ to resurrect 'classical cosmology' by applying the angular diameter-redshift test to compact radio sources. These sources are much smaller than the extended radio sources discussed in previous analyses, and Kellerman argues that one might therefore expect them to be less influenced by, for example, the evolution of the cosmological density. Kellerman indeed finds a minimum in the angular-size versus distance relationship corresponding to a high value of Ω_0 . A careful look at Kellerman's data, however, shows that he has not really detected a significant increase beyond the minimum. A model fit to his data in which the angular diameter decreases and then stays constant actually fits just as well as the $\Omega_0 = 1$ model he favours. Because the evolution of radio sources—which are highly dynamic and active objects—is likely to be extremely complicated, it may be that Kellerman has simply discovered a different form of time evolution than that found for extended radio sources. Until it is clearly demonstrated that this is not the case, it would be unwise to place too much emphasis on this claim, particularly as this is the only such analysis which has produced a high value of Ω_0 .

Number-counts have also been used to test cosmological geometry, but the results remain inconclusive³¹, so we will not discuss them here.

Gravitational lensing. In principle, the distortion of light paths due to the gravitational field of astronomical bodies provides a way of probing the cosmological density of material in a much more direct way than through number-counts or angular sizes¹.

One can either use the lensing of background objects to map the total amount of matter in a particular object, or use statistical arguments about the frequency of lensing of background sources to constrain the large-scale distribution of mass. The discovery of arcs and arclets in images of rich clusters of galaxies³² has led to estimates of their masses and density profiles which are not in contradiction with virial analysis methods described below, so these confirm the usual estimate of $\Omega_0 \approx 0.1-0.3$. It is in principle possible to extend this method to larger cluster radii by looking at the distortion of images of galaxies lying behind the cluster³³. There are indeed suggestions here that there is more dark matter at large radii than one would expect from virial arguments, but this is a relatively new technique and its uncertainties have not been fully assessed.

Large-scale structure

An outstanding problem in cosmology is the origin of galaxies and large-scale structure in the matter distribution. The standard model of structure formation involves the assumption that galaxies and clusters of galaxies grow by gravitational instability from small primordial perturbations in the energy density of the Universe. The form of these primordial perturbations to a large extent determines the nature of structure formed by the present epoch, but the rate at which clustering develops on different scales is also affected by cosmological factors, including the value of Ω_0 . One can attempt, therefore, to determine Ω_0 from clustering observations, but this is bound to be a model-dependent business.

The inflationary Universe suggests a possible origin for primordial density perturbations, in vacuum fluctuations of the quantum field responsible for driving inflation^{11,13}, but this origin is not obligatory. The standard model further assumes that the primordial inhomogeneities constitute a gaussian random field, that is, a stochastic superposition of spatial Fourier modes with random phases. The amplitude δ_k of each of the Fourier modes is a random number with a variance determined by the power spectrum, $P(k) \propto \langle \delta_k^2 \rangle$. In the standard version of the theory—which is predicted by simple models of inflation—the power spectrum is simply proportional to k (this is the Harrison-Zel'dovich spectrum); more generally, one might take $P(k) \propto k^n$, where n is the spectral index. During the expansion of the universe, the primordial spectrum is modified; one takes account of this by writing the power spectrum as $P'(k) = P(k)T^2(k)$ where $T(k)$, the transfer function, depends on the form of the dark matter and the time of matter-radiation equivalence, t_{eq} ; it therefore contains a dependence on Ω_0 which one might seek to exploit. In the CDM model (cold dark matter: dark matter consisting of particles which are non-relativistic when they decouple from the radiation background, for example, axions, photinos), a characteristic scale is imposed by the Hubble radius at t_{eq} , which scales as $(\Omega_0 h^2)^{-1}$. On the other hand, the relevant scale in the HDM model (hot dark matter: particles which are moving relativistically when they decouple, for example light massive neutrinos) is the neutrino free-streaming length which depends on the neutrino mass; and for models with only baryons, the relevant scale is the Silk damping scale.

To test this type of theory, one acquires a sample of galaxy redshifts from which one can infer some statistical description of its clustering properties. One can then compare this with either simulated samples or analytical calculations assuming some power spectrum. The main problem with this is that theory usually only deals with the distribution of gravitating material and does not provide an unambiguous way to specify sites where star formation is efficient enough to produce galaxies. This permits the existence of biased galaxy formation: the statistical properties of galaxies and mass fluctuations may not be directly comparable. The bias parameter b , defined as a multiplier between the fluctuations in total density $\delta\rho/\rho$ and the

fluctuations in counts of galaxies $\Delta N/N$, describes this:

$$\left(\frac{\delta N}{N}\right)_{\text{galaxies}} = b \left(\frac{\delta \rho}{\rho}\right)_{\text{matter}} \quad (5)$$

Usually b is assumed to be independent of scale. If this is true then the observed power spectrum of galaxy fluctuations is related to that of the mass fluctuations by $P_N(k) = b^2 P(k)$. Various astrophysical mechanisms have been discussed which might lead to some form of bias, but none has yet been definitely shown to be relevant to galaxy formation in any particular cosmological scenario³⁴⁻³⁸. However, if galaxy formation is indeed biased—and it would be surprising if it were not—then it is very doubtful whether it can be represented as a simple constant multiplier, except on very large scales³⁹, and perhaps not even then^{40,41}. Even the popular ‘high-peak bias’ model does not obey the relationship (5) with constant b .

Galaxy clustering. As mentioned above, one might aim to estimate Ω_0 by simply estimating the power spectrum of galaxy clustering and relating that to some assumed model such as CDM. If one assumes an $n=1$ primordial spectrum, one can determine Ω_0 by matching the observed shape with that expected for the transfer function appropriate for a given value of Ω_0 . The results of a number of independent studies⁴²⁻⁴⁸ demonstrate a power spectrum that has more large-scale power than CDM with $\Omega_0=1$, but which is well fitted by a CDM model with $\Omega_0 \approx 0.2$, as can be seen from Fig. 1, which shows a compilation of some of the latest power-spectrum results.

Galaxy random velocities. The weakness with simple galaxy clustering arguments is that they are misleading if there is a significant bias. One needs to study the distribution of all the mass, not just that part of it which happens to light up and form a galaxy. One way to do this is to look not just at the spatial positions of galaxies, but also at their peculiar velocities. (Peculiar velocity is the term given to deviations of galaxy motions from the pure ‘Hubble flow’: the radial velocity of a galaxy will, in general, be $v = H_0 r + v_p$ where v_p is the peculiar velocity.)

There are various ways¹ to use the statistical properties of peculiar motions in estimating Ω_0 . The anisotropy in statistical measures like the correlation function and power spectrum can

be used to estimate the magnitude of the radial component of the typical galaxy peculiar velocity. The typical galaxy relative peculiar velocities obtained by such methods are around 300 km s^{-1} . This information can be used to infer the total amount of mass using the statistical mechanics of self-gravitating systems. These methods consistently⁴⁹⁻⁵³ give estimates of Ω_0 in the range 0.1–0.3. If one allows a bias, however, these results can be made compatible with higher values of Ω_0 . Applications of such techniques to large-scale data, where one might hope to use simple linear theory, seem to yield slightly higher estimates of Ω_0 but with much larger errors⁵⁴. These estimates also agree with virial analyses of rich clusters of galaxies, in which the analysis is considerably simplified by assuming that the clusters are fully-relaxed and gravitationally-bound systems⁵⁵.

This value is about an order of magnitude larger than estimates of Ω_0 based on the mass-to-light ratios inferred for galaxy interiors^{1,56}. This discrepancy between galactic and extragalactic dark matter may be lessened by taking into account the fact that the dark component in galaxies is more dominant in less-luminous galaxies which are more numerous^{57,58}. Nevertheless, the amount of dark matter for which there is compelling direct evidence is a long way short of closing the Universe.

Bulk flows and streaming motions. A clearer picture of cosmic dynamics might be obtained by using peculiar motions, not in a statistical sense, but by trying to map out the flow pattern by directly measuring the peculiar velocity of each galaxy. The problem is that one must measure the galaxy distance using some indicator⁵⁹ other than its redshift. Knowing the distance, the contribution of a pure Hubble expansion at that distance can be subtracted to yield the galaxy’s peculiar velocity. If peculiar velocities are generated by inhomogeneities in the local distribution of material, they can be used to estimate the total amount of mass much as one can estimate the mass of the sun from the Earth’s orbital velocity.

The most impressive implementation of this idea is a method called POTENT⁶⁰⁻⁶². Rather than concentrating on statistical measures of clustering and velocity, POTENT attempts to construct a map of the density field in space using information about galaxy peculiar motions. This might seem impossible because one can measure only one component—the radial component—of the peculiar motion of each galaxy, and not the full three-dimensional velocity. However, there is an important property of gravitationally induced motions that overcomes this difficulty: in linear perturbation theory the peculiar velocity field is irrotational, and can therefore be expressed as the gradient of a potential function, say $\mathbf{v} = \nabla \Phi$. Even though we only know radial velocities, we can extract the total velocity potential Φ by simply integrating the radial velocity v_r along radial paths from the observer:

$$\Phi(r, \theta, \phi) = - \int_0^r v_r(r', \theta, \phi) dr' \quad (6)$$

By differentiating the velocity potential, one can recover the full three-dimensional velocity flow from the original one-dimensional data. From this velocity field, one can infer the distribution of mass that caused it, and comparing this inferred density field with the distribution of galaxies allows both Ω_0 and b to be constrained. In fact, in this and other linear peculiar motion studies, what usually emerges is a constraint on the combination $\Omega_0^{0.6}/b$.

The POTENT analysis, and other analysis of bulk flows⁶³, suggest a relatively high value of $\Omega_0^{0.6}/b$; these estimates are widely advertised as strong evidence for a critical-density universe. While it may well be the case that the POTENT results are consistent with a critical-density universe, it is less clear that they exclude the alternative, low-density case. Systematic errors might interfere in this analysis in various ways. To take an extreme example, suppose there is a systematic component within the error in estimated distances to galaxies that depends

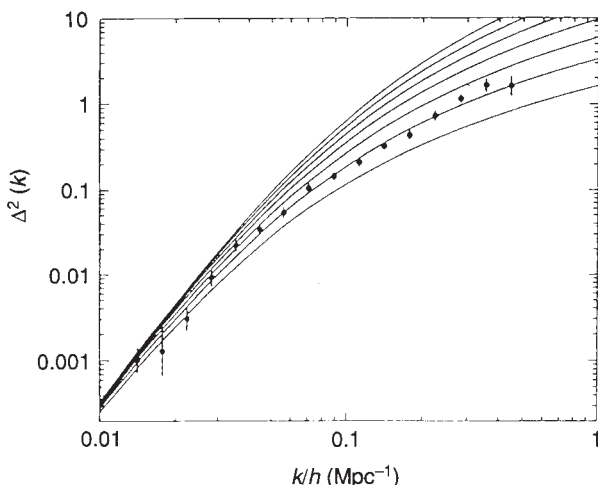


FIG. 1 Comparison of the ‘linearized’ power spectrum of galaxy clustering with various CDM models with different values of $\Omega_0 h$ (from ref. 48). The y-axis shows $\Delta^2(k) = k^3 P(k)$ as a function of wavenumber k . The data points are from a compilation of redshift surveys and the solid lines are various CDM models. The CDM models have $\Omega_0 h$ values of 0.5, 0.45, 0.25, 0.2. The best fit has $\Omega_0 h \approx 0.25$. (The linearization procedure does, in fact assume $\Omega_0 = 1$, but this procedure has only a small effect on large scales, where it is more likely that purely gravitational clustering is being recovered.)

on the local density. In this case the 'observed' peculiar velocities can be dominated by distance errors induced by density fluctuations, a potentially serious problem for POTENT⁶⁴. Intriguingly, a recent study by Guzman and Lucey⁶⁵ has shown that there is indeed such a systematic error in the most commonly used distance indicator for elliptical galaxies. What happens is that old stellar populations produce a different response in the distance indicator from young ones. Because older galaxies formed earlier and in higher-density environments, the upshot is exactly the short of systematic effect that is so dangerous to methods like POTENT. On the other hand, if the peculiar velocity data used in the largest POTENT analyses were dominated by such systematic errors, one would not expect to see such a good agreement as one does between the observed spatial distribution of galaxies and the inferred POTENT density field. It seems unlikely therefore that the most straightforward kind of systematic artefact suggested by Silk⁶⁴ is responsible for the POTENT maps. Nevertheless, until all possible systematic effects and Malmquist biases have been fully clarified, the significance of the POTENT result must remain an open question.

Dipole analysis. The CMB radiation has a dipole anisotropy corresponding to the motion of the Earth through the reference frame in which the radiation is isotropic. After subtracting the Earth's motion around the Sun, the Sun's motion around the Galactic Centre and the velocity of our Galaxy with respect to the centroid of the Local Group, this dipole anisotropy tells us the velocity of the Local Group through the cosmic reference frame. The result is a velocity of 600 km s^{-1} in the direction of Hydra-Centaurus⁶⁶ ($l = 268^\circ$, $b = 27^\circ$).

The gravitational instability picture explains this velocity as being due to the net gravitational pull on the Local Group through the inhomogeneous distribution of matter around it. If $\delta\rho/\rho$ can be estimated from a sufficiently large sample of galaxies $\Omega_0^{0.6}/b$ can in principle be determined, as long as the bias factor b is constant.

Suppose we have a sample of galaxies with some well-defined selection criterion so that the selection function—the probability that a galaxy at distance r from the observer is included in the catalogue—has some known form $\phi(r)$. Then the gravitational acceleration $\mathbf{g}$ can be approximated by

$$\mathbf{g} = \frac{4\pi G}{3} \mathbf{D} = G \sum_i \frac{1}{\phi(r_i)} \frac{\mathbf{r}_i}{r_i^2} \quad (7)$$

where the $\mathbf{r}_i$ are the galaxy positions and the sum is taken over all the galaxies in the sample. The dipole vector $\mathbf{D}$ can be calculated from the catalogue and, as long as it is aligned with the CMB dipole anisotropy, $\Omega_0^{0.6}$ can be estimated. However, note that this method measures only the inhomogeneous component of the gravitational field; it will not detect a mass component that is uniform on a scale larger than the sample size.

This technique has recently been very popular with cosmologists, mainly because the various IRAS galaxy catalogues are very suitable for this type of analysis^{67,68}. It does, however, have some difficulties, the most important being the problem of convergence. Suppose one has a catalogue that samples a small sphere around the Local Group, but this sphere is itself moving in the same direction (or similar) as the Local Group. For this to happen, the Universe must be significantly inhomogeneous on scales larger than the catalogue can probe. In this circumstance, the actual velocity explained by the dipole of the catalogue is not the whole CMB dipole velocity, but only a part of it. It follows, then, the $\Omega_0^{0.6}$ factor might be overestimated by comparing too large a velocity with the observed $\mathbf{D}$. One must be absolutely sure, therefore, that the sample is deep enough to sample all contributions to the Local Group motion. Using the deep QDOT redshift survey, Rowan-Robinson *et al.*⁶⁸ have demonstrated that there are significant contributions to the dipole from distances up to at least $100h^{-1} \text{ Mpc}$. However, using these data alone, we cannot know whether the dipole has actually converged to its final value within the limits of this sample.

So can one tell if there is indeed a significant contribution to $\mathbf{D}$ from large scales? Clearly, we need an object to trace the mass which does not have as steep a selection function as do IRAS galaxies. Rich clusters of galaxies fit the bill nicely; suitably chosen samples have selection functions which fall relatively slowly even beyond the QDOT depth. Analysis of cluster samples^{69,70} shows very clearly that there is a large contribution to the dipole from scales as large as $150h^{-1} \text{ Mpc}$. There is also circumstantial evidence for similar behaviour in the QDOT dipole itself⁷¹: although the QDOT sample is very sparse at $100h^{-1} \text{ Mpc}$, near the limit at which galaxies are bright enough to be in the catalogue, the dipole induced by a shell of galaxies around $150h^{-1} \text{ Mpc}$ is aligned within 13.6° of the CMB dipole, indicating strongly that there is a coherent anisotropy on this scale but that the sampling is too poor for this to be registered in the QDOT dipole (compare Fig. 2). The cluster data show a similar effect at $150h^{-1} \text{ Mpc}$, only stronger—alignment is within 6.4° , but because the sampling is much better for the clusters, there is a discernible increase in the cluster dipole on these scales (Fig. 2).

Contrary to previous claims^{67,68}, the QDOT dipole, interpreted in the light of the cluster data, strongly suggests a value of Ω_0 in the range 0.3–0.4. Note also that measurements of the dipole of the distribution of optical galaxies⁷² also lead to a low value of Ω_0 . Of course, there may be systematic problems with the cluster catalogues and clusters (or optical galaxies) may be highly biased objects, scepticism must be used for the cluster

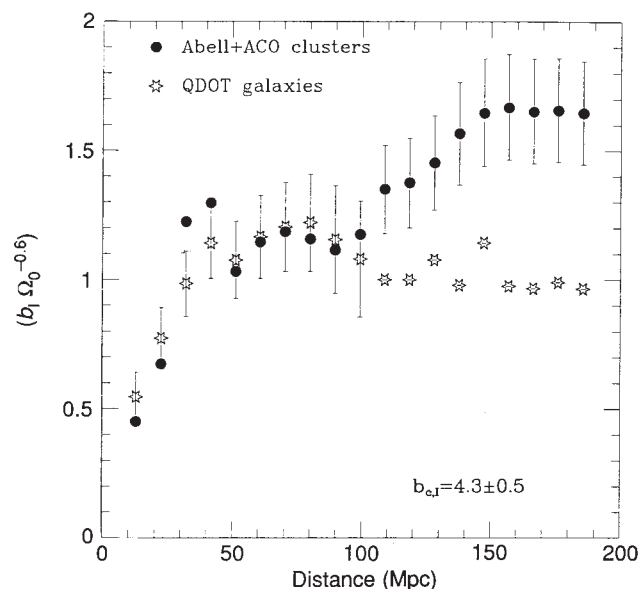


FIG. 2 A comparison of the dipole amplitude seen in Abell and ACO clusters, and QDOT galaxies (from M. Plionis, manuscript in preparation). The same binning was used for both samples, but the cluster dipole on small scales involves a correction for the Virgo and Puppis clusters, which is done by hand. The two dipoles were scaled on to the same axes by assuming a bias factor of clusters relative to IRAS galaxies of 4.3 ± 0.5 , as suggested by ref. 48. The y-axis is labelled in units of $b_l \Omega_0^{0.6}$, where b_l is the bias parameter of IRAS galaxies relative to mass. The final dipole vectors line up within 8° (clusters) and 20° (QDOT) from the CMB dipole. The agreement up to $100h^{-1} \text{ Mpc}$ is remarkable, but the cluster dipole has an additional contribution not seen in QDOT because of poor sampling; the QDOT dipole at $150h^{-1} \text{ Mpc}$ does, however, display a small 'blip' at $\sim 150h^{-1} \text{ Mpc}$, and a narrow shell at this distance has a dipole vector strongly aligned with the CMB dipole. Error bars are shown on half the points (for clarity): up to $100h^{-1} \text{ Mpc}$ we show the QDOT shot-noise errors calculated in ref. 71; the cluster dipole errors beyond $100h^{-1} \text{ Mpc}$ also include the uncertainty in the cluster bias factor.

data also. Moreover, recent work by Lauer and Postman⁷³ suggests that the large-scale cluster distribution may be executing a coherent motion itself; it is unclear what these results may imply for the density parameter. We should also point out that dipole analysis can be seriously affected by distortions if only redshift information is used to estimate distance⁷⁴.

The microwave background

In standard models of structure formation, small primordial perturbations to the energy density are amplified by gravitational instability as the Universe expands. Although these fluctuations were small at the time of recombination, $t = t_{\text{rec}} \approx 3 \times 10^5$ yr, they are predicted to leave a detectable imprint in the CMB radiation^{75,76}. On large angular scales (greater than a few degrees), the dominant mechanism by which density fluctuations induce anisotropy in the CMB is the Sachs–Wolfe effect: CMB photons suffer a gravitational redshift as they climb out of potential wells generated by overdensities at t_{rec} or a blueshift if they originate in density ‘troughs’. The COBE experiment⁷⁷, recently confirmed by an experiment on Tenerife⁷⁸, has indeed detected fluctuations in the temperature of the CMB on the sky of roughly the right amplitude to be consistent with this standard picture⁷⁹. What do the COBE results mean for Ω_0 ?

At early times the linear growth of fluctuations in an open universe is indistinguishable from the Einstein–de Sitter case. At later times, however, the expansion becomes curvature-dominated and, under these circumstances, density fluctuations can no longer grow at all. Roughly speaking, this occurs at a redshift $z_* \approx 1/\Omega_0 - 1$. To obtain a given level of fluctuation on a particular length scale at the present time ($z=0$) one requires that density fluctuations at recombination on that scale be a factor $\sim 1/\Omega_0$ higher than in a flat universe. This does not mean, however, that the CMB fluctuations on a given angular scale are predicted to be larger by the same amount. Two further effects intervene.

First, the usual Sachs–Wolfe relation between density inhomogeneities at last scattering and the CMB temperature anisotropy is calculated in a critical-density universe, where matter dominates the dynamics at the time of decoupling. In a low-density universe, however, pressure effects are significant at that time and alter the relation. By integrating the conservation equations along the null cone⁸⁰, it can be shown that there can be a considerable suppression of the amplitude of Sachs–Wolfe anisotropy in an open universe compared with the standard calculation.

Second, in a low-density universe, the relation between observational angles and corresponding linear scales in the last scattering surface is also altered significantly from that holding in a critical-density universe. As discussed above, structure formation requires the existence of a spectrum of fluctuations on different length scales. A given experiment, with a given angular resolution, will probe different parts of the power spectrum in an open universe or a flat universe because the length scale corresponding to a fixed angular scale is different in the two cases⁸¹. In fact, the angle θ subtended by a co-moving length scale l_0 Mpc is given by the approximate relation

$$\theta(l_0) \approx \frac{\Omega_0}{2c} l_0 H_0 \quad (8)$$

so an observation at a given angular scale provides a constraint on the fluctuations on a larger length scale in the open case. The net effect on the CMB anisotropy of lowering Ω_0 therefore depends on the relative amplitudes of fluctuations on different length scales, that is, on the power spectrum of the initial perturbations. For the usual Harrison–Zel’dovich spectrum of initial fluctuations (which is sometimes predicted by inflation), the net result is usually a lower CMB anisotropy than in the critical-density cases, on larger angular scales⁸² even though the fluctuations grow at a slower rate in open models. It is quite possible, therefore, to find open models compatible with the COBE results^{83–86}.

What is the verdict?

We have looked here at the main aspects of cosmology in which the density of the Universe plays a role as either a prediction or a parameter. We hope we have convinced the reader that no strongly convincing case can be made for a critical-density Universe, and that on the balance of the evidence, an open Universe should be preferred. We are not arguing that we have proved the Universe to be open—there is simply not sufficient evidence to make this claim—but we do argue that there are considerable grounds for doubting the assertion that we live in a critical-density Universe.

The central issue is what should be regarded as the ‘standard model’ of cosmology. What does it include? Clearly, there is agreement that it includes (1) expansion of the Universe, (2) from a hot Big Bang, (3) where nucleosynthesis of the light elements took place, (4) resulting in the CMB as relic radiation⁸⁷. Some workers would also like to include as part of the standard model the assumptions of (5) inflation in the past, (6) leading to a critical density today, and (7) with primaeval quantum fluctuations providing the seeds from which galaxy formation took place by gravitational instability of cold dark matter (or a mixture of hot and cold dark matter). We argue first that while the statements (1–4) are essential components of the standard model, the latter properties (5–7), exciting as they are, should not be regarded so, but rather as major contenders to become part of the standard model in the future when they have been convincingly confirmed experimentally. This suggestion would save the cosmological community from some of the embarrassment caused by misleading articles recently suggesting that COBE and/or recent large-scale structure data have undermined our standard understanding of cosmology, when all it has done is to show that the simplest models satisfying statements (5–7) are unviable.

Second, we argue that whether the elements (5–7) are regarded as part of the standard model or not, they should be tested by all possible observations in an open-minded way, just as the more basic elements (1–4) should. In particular, we argue that Ω_0 is an essential parameter in describing the standard models. As such, it should be determined by fitting to the data, together with other parameters such as the Hubble parameter H_0 , the bias parameter b and the primordial spectral index n . As we have shown, there is enough leeway in the present observations to allow a low-density model. Those cosmologists who take it for granted that we live in a high-density Universe—and there seem to be many—may turn out to be profoundly mistaken. □

Peter Coles is at the Astronomy Unit, School of Mathematical Sciences, Queen Mary and Westfield College, Mile End Road, London E1 4NS, UK. George Ellis is at the Department of Applied Mathematics, University of Cape Town, 7700 Rondebosch, Republic of South Africa.

1. Peebles, P. J. E. *Principles of Physical Cosmology* (Princeton Univ. Press, USA, 1993).
2. Weinberg, S. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972).
3. Peebles, P. J. E. *Nature* **321**, 27–32 (1986).
4. Cho, H. T. & Kantowski, R. A Measure on a Subspace of FRW Solutions and the ‘‘Flatness Problem’’ of Standard Cosmology. Preprint (Department of Physics and Astronomy, University of Oklahoma, Norman, 1994).
5. Gibbons, G. W., Hawking, S. W. & Stewart, J. M. *Nucl. Phys. B* **281**, 736–751 (1987).
6. Hawking, S. W. & Page, D. N. *Nucl. Phys. B* **298**, 789–809 (1988).
7. Ellis, G. F. R. in *Gravitation, Proc. Banff Summer Research Institute on Gravitation* (eds Mann, R. & Wesson, P.) 3–53 (World Scientific, Singapore, 1991).
8. Barrow, J. D. & Tipler, F. J. *The Anthropic Cosmological Principle* (Oxford Univ. Press, UK, 1986).
9. Garrett, A. J. M. & Coles, P. *Comments Astrophys.* **17**, 23–47 (1993).
10. Guth, A. H. *Phys. Rev. D* **23**, 347–356 (1981).
11. Kolb, E. & Turner, M. S. *The Early Universe* (Addison Wesley, Redwood, 1990).
12. Linde, A. *Inflation and Quantum Cosmology* (Academic, Boston, 1990).
13. Olive, D. *Phys. Rep.* **190**, 307–403 (1990).
14. Lucchin, F. & Matarrese, S. *Phys. Lett.* **164B**, 282–286 (1985).
15. Ellis, G. F. R. *Class. Qu. Grav.* **5**, 891–901 (1988).
16. Hubner, P. & Ehlers, J. *Class. Qu. Grav.* **8**, 333–346 (1991).
17. Ellis, G. F. R., Lyth, D. H. & Mijic, M. B. *Phys. Lett. B* **271**, 52–55 (1991).
18. Madsen, M. S. & Ellis, G. F. R. *Mon. Not. R. astr. Soc.* **234**, 67–77 (1988).
19. Sandage, A. R. *Astrophys. J.* **252**, 553–573 (1982).
20. Sybalsky, E. M. D. & Schramm, D. N. *Rep. Prog. Phys.* **44**, 293–328 (1981).
21. Carney, B. W. *Highlights of Astronomy* Vol. 6 (ed. West, R. M.) 255–266 (Reidel, Dordrecht, 1983).

22. Cannon, R. Q. *Highlights of Astronomy* Vol. 6 (ed. West, R. M.) 109–117 (Reidel, Dordrecht, 1983).
23. Rowan-Robinson, M. *The Cosmological Distance Ladder* (Freeman, New York, 1985).
24. Fukugita, M., Hogan, C. J. & Peebles, P. J. E. *Nature* **366**, 309–312 (1993).
25. Walker, T. P. et al. *Astrophys. J.* **376**, 51–69 (1991).
26. Gott, J. R., Gunn, J. E., Schramm, D. N. & Tinsley, B. M. *Astrophys. J.* **194**, 543–553 (1974).
27. White, S. D. M., Navarro, J. F., Evvard, A. E. & Frenk, C. S. *Nature* **366**, 429–433 (1993).
28. Hogan, C. J. *Ann. N. Y. Acad. Sci.* **647**, 76–85 (1991).
29. Sandage, A. R. *Phys. Today* **23**, 34–43 (1967).
30. Kellerman, K. I. *Nature* **361**, 134–136 (1993).
31. Fahlman, G., Kaiser, N., Squires, G. & Woods, D. *Astrophys. J.* (in the press).
32. Tyson, J. A., Valdes, F. & Wenk, R. A. *Astrophys. J.* **349**, L1–L4 (1990).
33. Gardner, J. P., Cowie, L. L. & Wainscoat, R. J. *Astrophys. J.* **415**, 9–12 (1993).
34. Kaiser, N. *Astrophys. J.* **284**, L9–L12 (1984).
35. Rees, M. J. *Mon. Not. R. astr. Soc.* **213**, 75P–81P (1985).
36. Dekel, A. *Comments Astrophys.* **11**, 235–256 (1986).
37. Bardeen, J. M., Bond, J. R., Kaiser, N. & Szalay, A. S. *Astrophys. J.* **304**, 15–61 (1986).
38. Dekel, A. & Rees, M. J. *Nature* **326**, 455–462 (1987).
39. Coles, P. *Mon. Not. R. astr. Soc.* **262**, 1065–1075 (1993).
40. Babul, A. & White, S. D. M. *Mon. Not. R. astr. Soc.* **253**, 31P–34P (1991).
41. Bower, R. G., Coles, P., Frenk, C. S. & White, S. D. M. *Astrophys. J.* **405**, 403–412 (1993).
42. Maddox, S. J., Efstathiou, G., Sutherland, W. J. & Loveday, J. A. *Mon. Not. R. astr. Soc.* **242**, 43P–47P (1990).
43. Collins, C. A., Nichol, R. C. & Lumsden, S. L. *Mon. Not. R. astr. Soc.* **254**, 295–300 (1992).
44. Efstathiou, G., Sutherland, W. J. & Maddox, S. J. *Nature* **348**, 705–707 (1990).
45. Efstathiou, G. et al. *Mon. Not. R. astr. Soc.* **247**, 10P–14P (1990).
46. Saunders, W. et al. *Nature* **349**, 32–38 (1991).
47. Peacock, J. A. *Mon. Not. R. astr. Soc.* **253**, 1P–4P (1991).
48. Peacock, J. A. & Dodds, S. J. *Mon. Not. R. astr. Soc.* **267**, 1020–1034 (1994).
49. Davis, M. & Peebles, P. J. E. *Astrophys. J.* **267**, 465–482 (1983).
50. Bean, A. J., Efstathiou, G., Ellis, R. S., Peterson, B. A. & Shanks, T. *Mon. Not. R. astr. Soc.* **205**, 605–624 (1983).
51. Peebles, P. J. E. *Science* **224**, 1385–1391 (1984).
52. Hale-Sutton, D., Fong, R., Metcalfe, N. & Shanks, T. *Mon. Not. R. astr. Soc.* **237**, 569–587 (1989).
53. Tully, R. B. *Astrophys. J.* **321**, 280–304 (1987).
54. Fisher, K. B., Davis, M., Strauss, M. A., Yahil, A. & Huchra, J. P. *Mon. Not. R. astr. Soc.* **267**, 927–948 (1994).
55. Jones, C. & Forman, W. *Astrophys. J.* **276**, 38–55 (1984).
56. Richstone, D. O. & Tremaine, S. *Astr. J.* **92**, 72–74 (1986).
57. Persic, M. & Salucci, P. *Mon. Not. R. astr. Soc.* **234**, 131–154 (1988).
58. Persic, M. & Salucci, P. *Mon. Not. R. astr. Soc.* **245**, 577–581 (1990).
59. Burstein, D. *Rep. Prog. Phys.* **53**, 421–481 (1990).
60. Bertschinger, E. & Dekel, A. *Astrophys. J.* **336**, L5–L8 (1989).
61. Dekel, A., Bertschinger, E. & Faber, S. M. *Astrophys. J.* **364**, 349–369 (1990).
62. Dekel, A. et al. *Astrophys. J.* **412**, 1–21 (1993).
63. Bertschinger, D., Dekel, A., Faber, S. M. & Burstein, D. *Astrophys. J.* **364**, 370–395 (1990).
64. Silk, J. *Astrophys. J.* **345**, L1–L4 (1989).
65. Guzman, R. & Lucey, J. R. *Mon. Not. R. astr. Soc.* **263**, L47–L50 (1993).
66. Lubin, P. M., Epstein, G. L. & Smoot, G. F. *Phys. Rev. Lett.* **50**, 616–619 (1983).
67. Strauss, M. & Davis, M. in *Proceedings of Vatican Study Week on Large-scale Motions in the Universe* (eds Rubin, V. C. & Coyne, G.) 256–277 (Princeton Univ. Press, USA, 1988).
68. Rowan-Robinson, M. et al. *Mon. Not. R. astr. Soc.* **247**, 1–18 (1990).
69. Scaramella, R., Vettolani, G. & Zamorani, G. *Astrophys. J.* **376**, L1–L4 (1991).
70. Plionis, M. & Valdarnini, R. *Mon. Not. R. astr. Soc.* **249**, 46–61 (1991).
71. Plionis, M., Coles, P. & Catelan, P. *Mon. Not. R. astr. Soc.* **262**, 465–474 (1993).
72. Hudson, M. J. *Mon. Not. R. astr. Soc.* **265**, 72–80 (1993).
73. Lauer, T. R. & Postman, M. *Astrophys. J.* **425**, 418–438 (1994).
74. Kaiser, N. *Mon. Not. R. astr. Soc.* **227**, 1–21 (1987).
75. Hogan, C. J., Kaiser, N. & Rees, M. J. *Phil. Trans. R. Soc. A* **307**, 97–109 (1982).
76. Kaiser, N. & Silk, J. *Nature* **324**, 529–537 (1987).
77. Smoot, G. F. et al. *Astrophys. J.* **396**, L1–L5 (1992).
78. Hancock, S. et al. *Nature* **367**, 333–338 (1994).
79. Wright, E. L. et al. *Astrophys. J.* **396**, L13–L18 (1992).
80. Stoeger, W. R., Ellis, G. F. R. & Xu, C. *Phys. Rev. D* **49**, 1845–1853 (1994).
81. Ellis, G. F. R. & Tavakol, R. *Class. Qu. Grav.* **11**, 675–688 (1994).
82. Traschen, J. E. & Eardley, D. M. *Phys. Rev. D* **34**, 1665–1679 (1986).
83. Holtzman, J. *Astrophys. J. Suppl. Ser.* **71**, 1–24 (1989).
84. Gouda, N., Sugiyama, N. & Sasaki, M. *Astrophys. J.* **372**, L49–L52 (1991).
85. Wilson, M. L. *Astrophys. J.* **273**, 2–15 (1983).
86. Kamioniski, M. & Spergel, D. N. *Astrophys. J.* (in the press).
87. Peebles, P. J. E., Schramm, D. N., Turner, E. L. & Kron, R. G. *Nature* **352**, 769–776 (1991).

ACKNOWLEDGEMENTS. We thank D. W. Sciama, P. Salucci, C. J. Hogan, C. J. Isham and, particularly, P. J. E. Peebles for comments; and M. Plionis for preparing Fig. 2 and for comments. P.C. received a SERC Advanced Research fellowship. We both thank the SERC for support under the QMW Visitor's Programme, and G.F.R.E. thanks the FRD (South Africa) for financial support.

Controlling chaos in the brain

Steven J. Schiff^{*}, Kristin Jerger^{*}, Duc H. Duong^{*}, Taeun Chang^{*},
Mark L. Spano[†] & William L. Ditto[‡]

^{*} Department of Neurosurgery, Children's National Medical Center
and The George Washington University School of Medicine, Washington DC 20010, USA

[†] Naval Surface Warfare Center, White Oak Laboratory, Silver Spring, Maryland 20903, USA

[‡] School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

In a spontaneously bursting neuronal network *in vitro*, chaos can be demonstrated by the presence of unstable fixed-point behaviour. Chaos control techniques can increase the periodicity of such neuronal population bursting behaviour. Periodic pacing is also effective in entraining such systems, although in a qualitatively different fashion. Using a strategy of anticontrol such systems can be made less periodic. These techniques may be applicable to *in vivo* epileptic foci.

FOLLOWING the recent theoretical prediction that chaotic physical systems might be controllable with small perturbations^{1,2}, there has been rapid and successful application of this technique to mechanical systems³, electrical circuits⁴, lasers⁵ and chemical reactions^{6,7}. Following the demonstration of the control of chaos in arrhythmic cardiac tissue⁸, there are no longer any technical barriers to applying these techniques to neural tissue.

One of the hallmarks of the human epileptic brain during periods of time in between seizures is the presence of brief bursts of focal neuronal activity known as interictal spikes. Often such spikes emanate from the same region of brain from which the seizures are generated but the relationship between the spike patterns and seizure onsets remains unclear^{9,10}. Several types

of *in vitro* brain slice preparations, usually after exposure to convulsant drugs that reduce neuronal inhibition, exhibit population burst-firing activity that in many ways seems analogous to the interictal spike¹¹. One of these preparations is the high potassium concentration ($[K^+]$) model, where slices from the hippocampus of the temporal lobe of the rat brain (a frequent site of epileptogenesis in the human) are exposed to artificial cerebrospinal fluid containing 6.5–10 mM $[K^+]$ ¹². After exposure to high $[K^+]$, spontaneous bursts of synchronized neuronal activity originate in a region known as the third part of the cornu ammonis or CA3 (ref. 13). Impulses from the CA3 bursts are propagated through a recurrent collateral fibre tract (the Schaffer collateral fibres) from CA3 to CA1, where electrographic seizure-like discharges can frequently be

observed¹⁴. A detailed computer model of the high $[K^+]$ burst discharges in CA3 successfully replicates many of the experimental findings¹⁵ (reviewed in ref. 16). Although it is difficult to identify determinism in long time series of such bursting activity using nonlinear prediction techniques, some evidence of determinism was recently identified^{17,18}. We sought to determine whether such neuronal bursting activity was amenable to control.

There have been substantial efforts to influence neuronal activity with electric fields and currents. Regarding the *in vitro* hippocampal slice, brief direct current (d.c.) currents from non-polarizable electrodes in the perfusion bath can influence the evoked excitability of pyramidal cells in normal $[K^+]$ when the electric fields are suitably oriented¹⁹. Apparently similar effects can be achieved with brief currents from monopolar polarizable microelectrodes placed directly into the tissue^{20,21}. To our knowledge the use of electrical stimulation to entrain spontaneous burst discharges from CA3 in high $[K^+]$ has not been attempted,

whether indirectly with electric fields or with direct stimulation of the nerve cells themselves. Direct stimulation can be accomplished for CA3 neurons by stimulating the Mossy fibre neurons that synapse on the CA3 pyramidal cells (orthodromic stimulation) or by stimulating branches of the CA3 pyramidal cell axons (Schaffer collateral fibres) and allowing the action potentials to propagate into the pyramidal cells in a retrograde fashion (antidromic stimulation).

With the observation that it was feasible to entrain burst discharges from CA3 with both orthodromic and antidromic stimulation (K.J. and S.J.S., manuscript in preparation), we were in a position to ask the following questions: (1) is there evidence for deterministic chaotic behaviour in this preparation; and (2) could such activity be controlled?

If one observes the timing of events from a chaotic physical system, those events are aperiodic. The timing of events evolves from one unstable periodicity to another. Furthermore the approach to these unstable periodicities shows recurring patterns

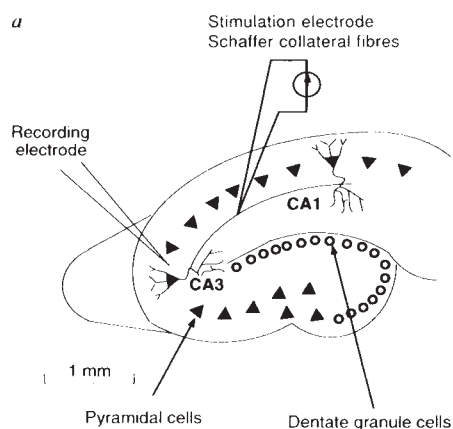
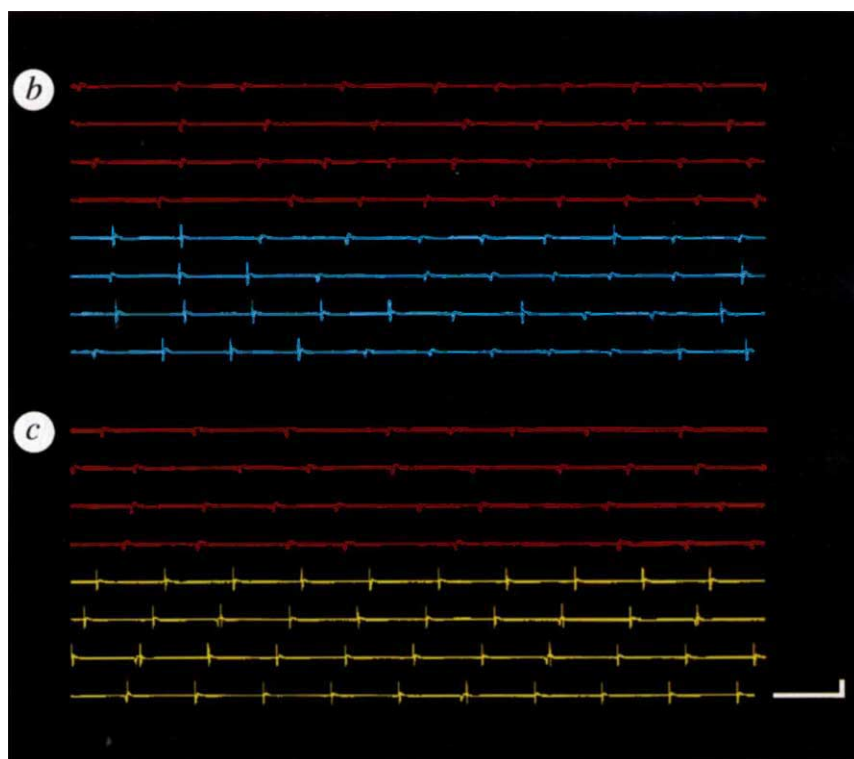
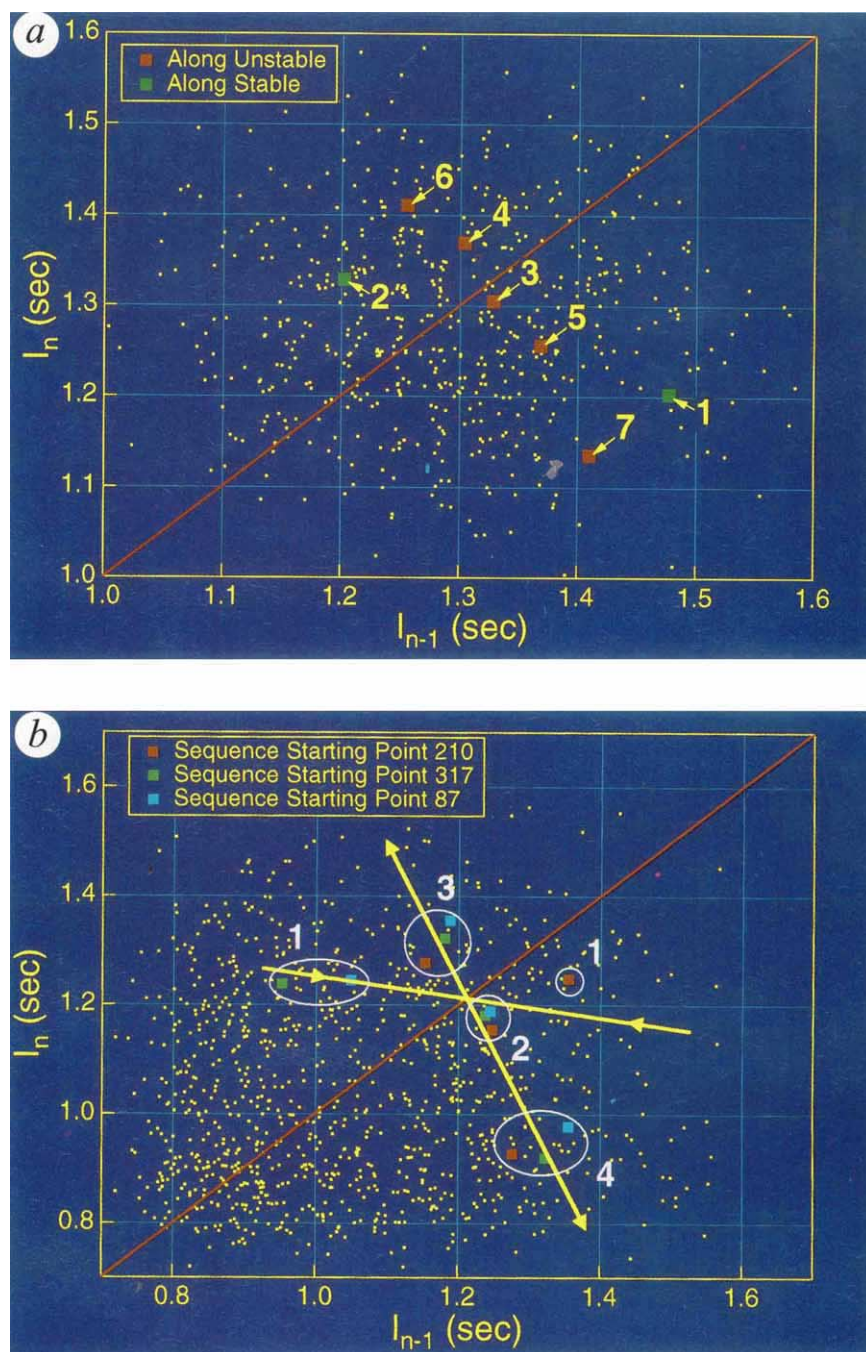


FIG. 1 a, Schematic diagram of the transverse hippocampal slice, and arrangement of recording electrodes. Female Sprague-Dawley rats weighing 125–150 g were anaesthetized with diethyl-ether and decapitated. Transverse slices 400 μ m thick were prepared from the hippocampus with a tissue chopper and placed in an interface-type perfusion chamber at 32–35 °C. Slices were perfused with artificial cerebrospinal fluid (ACSF) flowing at 2 ml min⁻¹ and composed of 155 mM Na⁺, 136 mM Cl⁻, 3.5 mM K⁺, 1.2 mM Ca²⁺, 1.2 mM Mg²⁺, 1.25 mM PO₄³⁻, 24 mM HCO₃⁻, 1.2 mM SO₄²⁻, and 10 mM dextrose. After 90 min of incubation, slices were tested for viability by recording a greater than 2 mV unitary population spike in the stratum pyramidale of CA1, in response to stimulation of Schaffer collateral fibres in the stratum radiatum with 100 μ s constant current 50–150 μ A square-wave pulses delivered at 0.1 Hz through tungsten microelectrodes. Recordings were made with 2–4 M Ω glass needle electrodes filled with 150 mM NaCl. With confirmation of viability, the perfusate was switched to ACSF containing 8.5 mM $[K^+]$ and 141 mM $[Cl^-]$. After 15–20 min of high $[K^+]$ perfusion, spontaneous burst firing could be recorded from CA3a or CA3b. Recordings were digitized across 12 bits at 5 kHz with a Digidata 1200 analogue to digital converter (Axon Instruments), and stored on a personal computer using Axotape 2.0 (Axon Instruments). CA3 interburst intervals were measured with Datapac II (Run Technologies). For control purposes, a separate computer system, digitizing across 16 bits at 1 kHz, identified spontaneous bursts from CA3 using a threshold and peak amplitude detection strategy. This computer triggered a stimulator (Model S8800, Grass Corp.) linked to a photoelectric stimulus isolation unit (Model SIU7, Grass Corp.), to deliver 100- μ s constant current square-wave pulses to the Schaffer collateral fibres through a tungsten microelectrode. At times, double pulses consisting of pairs of 100- μ s pulses with 150- μ s interpulse intervals were used. **b**, Shown are 100 s



of recording from an extracellular electrode within CA3 following exposure to 8.5 mM $[K^+]$. The upper four traces in red show the burst discharges occurring at irregular intervals. The lower four traces in blue show the effect of turning on chaos control. When control of chaos pulses are delivered to the Schaffer collateral fibres, large stimulation artefacts are seen at the onset of a burst. Note that each control pulse evokes a synchronized discharge in CA3. As control becomes more stable, the bursts become increasingly periodic, as seen most clearly on the bottom trace. Note that the control pulses are delivered only intermittently. The plot of the full series of these intervals before and during control is seen in the first half of the plot in Fig. 3. **c**, Shown are 100 s of recording from the same experiment before (red) and after (yellow) the initiation of periodic pacing. Note that periodic pulses immediately entrain the population bursts in the fifth tracing (10/10 pulses), but by the sixth tracing, there is occasional escape behaviour evident when the population bursts just before the delivered pulse (seen in 3/10 bursts in tracing 6, 2/11 in tracing 7, and 2/10 in tracing 8). The plot of the full series of these intervals before and during periodic pacing is seen in the second half of the plot in Fig. 3. The calibration mark indicates 5 mV ordinate and 1 s abscissa.

FIG. 2 a, Return plots of interburst intervals I_n versus the previous interval I_{n-1} without control. Seven sequential points are colour coded and numbered 1–7. Note that as the trajectory crosses the line of identity along a particular direction from 1 to 2, the next points take a peculiar sequence that starts close to the line of identity at point 3, and then alternates on either side of the line of identity and progressively diverges from it along a nearly straight line for points 3–7. The points coloured in green (1, 2), define a stable direction or manifold, whereas the points in red (3–7) define an unstable manifold. The intersection of these manifolds with the line of identity defines the unstable fixed point. Note that, as for other physical and biological systems^{3,8}, the saddle observed in these preparations is a flip saddle; that is, while the distances of successive state points from the fixed point increase in an exponential fashion along the unstable manifold, the state points alternate on opposite sides of the line of identity. b, Return plot showing multiple colour-coded trajectories that did not follow each other in time. The starting point for each sequence, numbered 1 in each, began at burst numbers 87 (blue), 210 (red) and 317 (green), out of a total series of 320 bursts. The stable manifold is shown with arrows pointing towards the unstable fixed point, and the unstable manifold has arrows drawn in the direction away from the unstable fixed point. Note that each sequence starts at a point roughly the same distance from the unstable fixed point in a region close to the stable manifold. Each trajectory follows roughly the same path, and after closely approaching the unstable fixed point for the points labelled 2 (red, green and blue), there evolves for each trajectory a sequence of exponentially diverging jumps (on alternating sides of the line of identity along the unstable manifold). Although the trajectories are clustered along the unstable manifold for four bursts, the fifth bursts are widely scattered (not shown). This is the manifestation of the sensitivity to initial conditions typical of chaotic systems. These plots show clear evidence for non-random structure in these trajectories, and their patterns bear the hallmarks of deterministic chaos.



which can be quantitatively understood by examining the relationship between the timing of sequential events. This can be visualized using a plot which is a type of a return map. Such a map plots the present interval between events, I_n , versus the j th previous interval, I_{n-j} . Periodicities (of period j) are revealed on such a plot as intersections with the line of identity, $I_n = I_{n-j}$. In a chaotic system these intersections are known as unstable fixed points. These unstable fixed points are deterministically approached from a direction called the stable direction or manifold and exponentially diverge from these points along the unstable direction or manifold. This type of local geometry has the shape of a saddle. Chaos control^{1,3,8} consists of the identification and characterization of these saddles, followed by a control intervention which exploits the local geometry of the saddle to increase the periodic behaviour of the system.

We demonstrate here three separate approaches to control: periodic pacing, an implementation of chaos control theory, and the inverse of chaos control which we term anticontrol.

Brain slice preparation

Hippocampal slices were prepared using standard techniques¹⁸. Glass micropipette electrodes in CA1 and CA3 were used to record neuronal activity, and a computer calculated and delivered appropriately timed control pulses. The anatomy of the transverse hippocampal slice and arrangement of electrodes are illustrated in Fig. 1a, and details of the experimental preparation are summarized in the figure legend.

Chaos identification and control

A key feature of chaotic systems is that they contain an infinite number of unstable periodic fixed points²². That these spontaneously active neuronal networks are chaotic was supported by evidence of unstable fixed points in the return maps with the characteristics shown in Fig. 2a, b. These candidate unstable fixed points met four criteria. First, the sequence of points approaches the unstable fixed-point candidate along a stable direction and diverges from it along an unstable direction.

FIG. 3 Plot showing chaotic (red) interburst intervals I_n before and after single-pulse chaos control (blue) and single-pulse periodic pacing (yellow). Burst number is indicated by n . Note the qualitative differences in the control achieved with each method. Raw data from the onsets of chaos control and periodic pacing are shown in Fig. 1b, c.

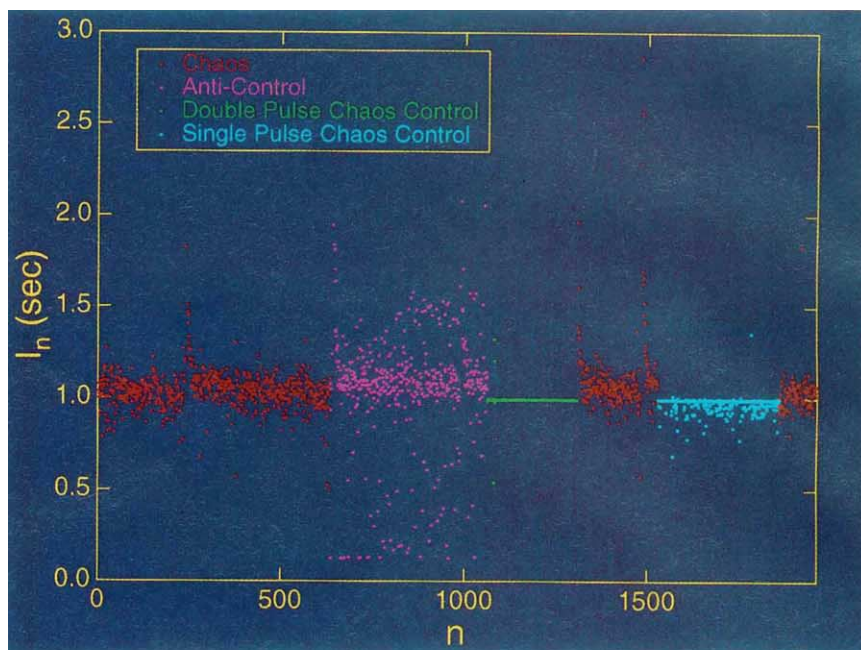
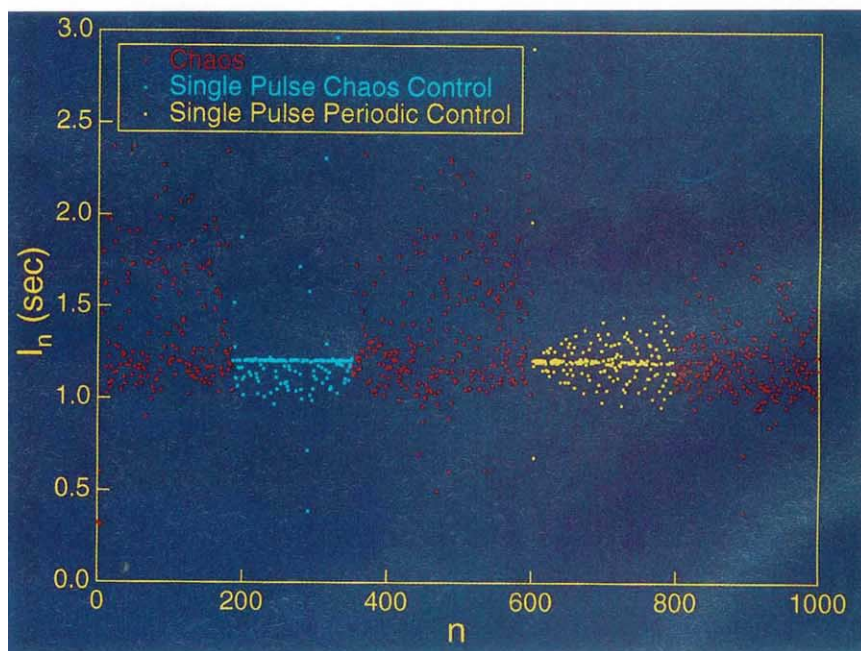


FIG. 4 Comparison of anticontrol (pink), double-pulse chaos control (green) and single-pulse control (blue). Note how anticontrol reduces the periodicity of the preparation. Also note that double-pulse control is more effective than single-pulse control, as was generally the case (see Table 1).



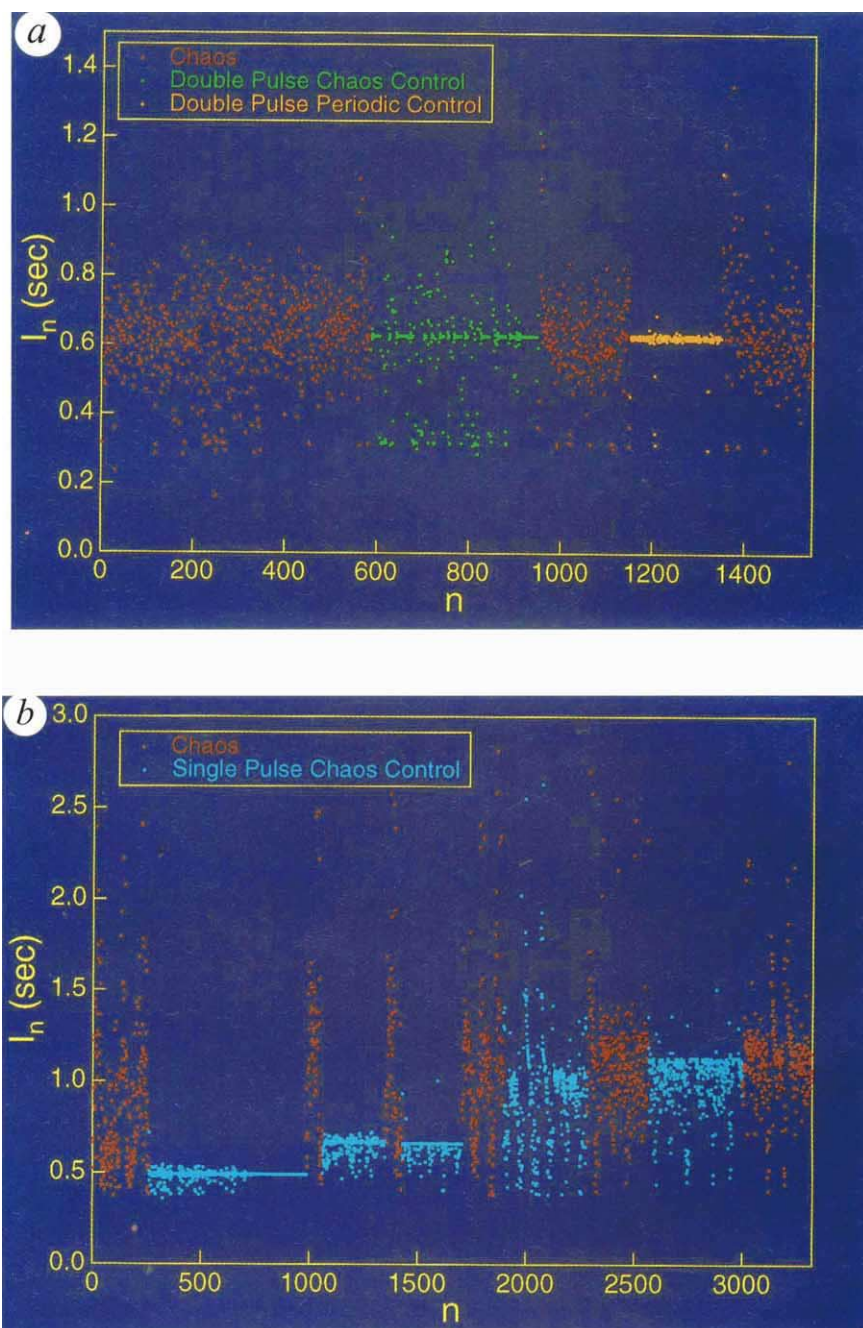
Second, the departing trajectory must be locally linear. Third, multiple approaches to the same candidate along the same stable direction with the corresponding departures along the same unstable direction must be detected. Last, the distances of the departing points from the candidate fixed point must increase exponentially, thereby demonstrating the sensitivity to initial conditions that is the defining feature of chaos. Although short single approaches to candidate unstable fixed points can be occasionally observed in random data, multiple approaches fitting these stringent criteria will not be seen.

In the implementation of these criteria, we excluded higher period behaviour (period 2, period 3, and so on), insisted that the geometry about the fixed point candidates be a flip saddle, and set criteria for the minimum acceptable linearity of the unstable manifold. We also set criteria for limits on the rates of approach and divergence of the trajectories. Because on these return plots it is easily shown that the eigenvalues (local rates of expansion and contraction along the respective manifolds)

are simply the slopes of the respective manifolds, it is important to note that we found that the stable manifolds had negative slopes with magnitudes less than 1 and that the unstable manifolds had negative slopes with magnitudes greater than 1, further supporting the presence of chaos in these data (see Fig. 2).

Our chaos control technique begins with a learning phase consisting of the identification of unstable fixed points and the performance of local linear least-square fits to obtain the stable and unstable directions along with the rates of approach and divergence along them. The control phase consisted of waiting until the system executed a close approach to the unstable fixed point (within a small radius ϵ) along the stable direction, followed by an intervention that modified the timing of the predicted next interval in order to place it back onto the stable manifold. In this way we use the saddle structure inherent in the chaotic dynamics to bring the system back to the desired unstable fixed point with minimal intervention.

FIG. 5 Comparison of double-pulse chaos control (green) with double-pulse periodic pacing (orange) in an exceptional experiment where double-pulse chaos control was not very tight, and where double pulse periodic pacing was a better control method. Note the frequent escapes and recaptures of control with the chaos control method. This example helps illustrate how the chaos control method differed from overdrive periodic pacing. *b*, Experimental run where a wide range of control behaviours were observed. There are 5 sequential attempts to identify unstable fixed points and learn the manifold structure. Each control run is coded in blue. The first control run demonstrates what we classify as good control in Table 1. After relearning, the second attempt at control was also good, but at a higher fixed point (where the period of the underlying periodic orbit of the fixed point is larger). The third trial, despite learning nearly the identical fixed point as the previous trial, chose different manifolds and the control was not as good. We call such control runs, where intervals I_n larger than the fixed point are selectively eliminated partial control; note that this is a manifestation of the manifold selection and not just a function of the mean frequency of stimulation. This point is further clarified in the fourth and fifth control runs. The fourth control run selected a larger fixed point than the previous runs but control was a failure. The fifth and final control run of this sequence, despite selection of the highest fixed point of this series, achieved partial control with more adequate manifold and fixed-point selection.



In the following experiments, for periodic pacing, we used the same pulse interval as our calculated unstable fixed point. For anticontrol, we chose an interval that placed the next point on a line that was completely off the manifolds. We somewhat arbitrarily chose a direction that was the mirror image of the unstable manifold about the line of identity. This anticontrol technique proved effective in diverting the state of the system away from the stable direction and therefore from the unstable fixed point.

Experimental results

We did 91 experimental trials on 22 hippocampal slices from 9 rats. These trials consisted of combinations of chaos control, periodic pacing and anticontrol using both single and double stimulation pulses (Table 1). Double pulses were used at times to increase the effectiveness of stimulation.

Figure 1*b* shows an example of the synchronized burst discharges recorded extracellularly in CA3. The upper four traces show the burst discharges occurring at irregular intervals before

control. On the fifth tracing chaos control pulses are delivered to the Schaffer collateral fibres, seen as large artefacts at the onset of a burst. Note that each control pulse can evoke a synchronized discharge in CA3, but control pulses are only given intermittently. As control becomes more stable, as seen most clearly on the bottom trace, the spontaneous bursts become increasingly periodic. We examined the relationship between burst intensity (duration, integrated amplitude and root mean

TABLE 1 Summary of experiments

	Chaos control		Periodic pacing		Anticontrol	
	Single	Double	Single	Double	Single	Double
Good	10	4	6	2	4	1
Partial	15	2	5	0	—	—
Bad	21	0	6	0	16	0
Total	46	6	17	2	20	1

square deviation) and interburst interval but found no correlations. The plot of the full series of these intervals before and during control is seen in the first half of the plot in Fig. 3.

Figure 1c shows 100 s of recording from the same experimental run before and after the initiation of periodic pacing. Note that now pulses immediately entrain the population bursts in the fifth tracing, but by the sixth tracing there is occasional escape behaviour evident when the population burst-fires just before the delivered pulses. The plot of the full series of these intervals before and during periodic pacing is seen in the second half of the plot in Fig. 3.

Figure 3 illustrates burst intervals before, during and after both single-pulse chaos control and single-pulse periodic pacing. Note that there is a qualitative difference in the control achieved with each method, as was seen from the traces of raw data shown in Fig. 1b, c.

Figure 4 compares the effects of anticontrol, double-pulse chaos control and single-pulse chaos control for another experiment. Here, the effect of anticontrol in reducing the periodicity of the preparation is clearly seen. We also demonstrated an increase in the effectiveness of double over single pulsing for this experiment.

Figure 5a illustrates an experiment where double-pulse chaos control and periodic pacing were compared, but the degree of control with chaos control was not as good as with periodic pacing. This experiment was exceptional, because most of the double-pulse experiments achieved very tight control (see Fig. 4, and summary in Table 1). Note that even with the increased stimulation of double pulsing, chaos control was not simply overdrive pacing, in that there were frequent escapes and recaptures of control consistent with the interactive control hypothesis.

Figure 5b illustrates the full gamut of results obtained with single-pulse chaos control. The first control trial demonstrated good control. Note that during this control sequence, the quality of the control improved. We then allowed the algorithm to relearn the fixed point and it chose one with a longer interval, again with good control. Whether this is a manifestation of the non-stationary nature of the system, or truly reflects the presence of multiple unstable period 1 fixed points is unclear. Relearning a third time, the algorithm selected different manifolds without substantially changing the value of the fixed point, but the control is not as good. This third control run exhibited what we call partial control in the results shown in Table 1; in partial control, manifold selection was not optimal, and time intervals longer than those of the chosen fixed-point interval appear eliminated. Relearning a fourth time, a fixed point is chosen at a longer interval than the previous three attempts. This was a control failure. Nevertheless, relearning a fifth time with an even longer

fixed-point interval gave partial control. This sequence demonstrates that the quality of control is not simply a function of the frequency chosen, the control is critically dependent on the quality of the fixed points and manifolds selected.

Discussion

This is the second attempt at achieving control of a chaotic biological system⁸ with a derivative method of Ott, Grebogi and Yorke¹. The observation of small-scale structure and the identification of stable and unstable manifolds near unstable fixed points for many of these burst-firing slices demonstrated the presence of deterministic chaos in this simple neuronal system. For this preparation, complicated control theory is not required just to achieve relatively fast periodic behaviour. Above certain frequencies, periodic pacing was effective in entraining the spontaneous burst discharges in CA3. However, the quality of control with periodic pacing was not equivalent with the chaos control method; furthermore, the chaos control method has the advantage over overdrive periodic pacing in terms of its ability to identify and track fixed points over time. In addition, the control of chaos strategy offers the ability to break up fixed-point behaviour with anticontrol. The anticontrol method used here uses a minimal number of stimuli needed to prevent periodic behaviour.

Although it was easy to observe the unstable manifolds from this preparation, it was difficult to place accurately the stable manifold. This is because the system has many degrees of freedom (that is, is high dimensional), and the expectation of fully disentangling its dynamics with a two-dimensional embedding is simplistic. In addition, increasing amounts of noise in a system will increase the frequency of escapes from control and eventually render control impossible¹. Despite these difficulties, good or partial control could frequently be achieved with our implementation of chaos control. The application of new theoretical techniques recently developed for the control of high dimensional systems²³ could improve the reliability of control for such neuronal preparations.

We experimented with several variants of stimulation delivery. In addition to single pulses, double pulses were used and were often more effective in achieving higher quality control. We also explored limiting the delivery of control pulses by prohibiting consecutive control pulses without an intervening spontaneous burst. This latter method was less effective than permitting consecutive pulsing.

Because this neuronal preparation shares similar characteristics with epileptic interictal spike foci, we believe these methods may be applied to such foci. Although it is impossible to predict what effect increasing the periodicity of epileptic foci will have, the opposite effect of breaking up fixed-point periodic behaviour with anticontrol could be a more useful intervention. □

Received 11 May; accepted 5 July 1994.

- Ott, E., Grebogi, C. & Yorke, J. A. *Phys. Rev. Lett.* **64**, 1196–1199 (1990).
- Shinbrot, T., Grebogi, C., Ott, E. & Yorke, J. A. *Nature* **363**, 411–417 (1993).
- Ditto, W. L., Raueo, S. N. & Spano, M. L. *Phys. Rev. Lett.* **65**, 3211–3214 (1990).
- Hunt, E. R. *Phys. Rev. Lett.* **67**, 1953–1955 (1991).
- Roy, R., Murphy, T. W., Maier, T. D. & Gills, Z. *Phys. Rev. Lett.* **68**, 1259–1262 (1992).
- Petrov, V., Gaspar, V., Masere, J. & Showalter, K. *Nature* **361**, 240–243 (1993).
- Rollins, R. W., Parmananda, P. & Sherard, P. *Phys. Rev. E* **47**, R780 (1993).
- Garfinkel, A., Spano, M., Ditto, W. L. & Weiss, J. *Science* **257**, 1230–1235 (1992).
- Gotman, J. *Can. J. Neurol. Sci.* **18**, 573–576 (1991).
- Katz, A., Marks, D. A., McCarthy, G. & Spencer, S. S. *Electroenceph. clin. Neurophysiol.* **79**, 153–156 (1991).
- Pedley, T. A. & Traub, R. D. in *Current Practice of Clinical Electroencephalography* 2nd edn (eds Daly, D. D. & Pedley, T. A.) 107–137 (Raven, New York, 1990).
- Rutecki, P. A., Lebeda, F. J. & Johnston, D. J. *Neurophysiol.* **54**, 1363–1374 (1985).
- Korn, S. J., Giacchino, J. L., Chamberlin, N. L. & Dingleline, R. J. *Neurophysiol.* **57**, 325–341 (1987).

- Traynelis, S. F. & Dingleline, R. J. *Neurophysiol.* **59**, 259–276 (1988).
- Traub, R. D. & Dingleline, R. J. *Neurophysiol.* **64**, 1009–1018 (1990).
- Traub, R. D. & Miles, R. *Neuronal Networks of the Hippocampus* (Cambridge Univ. Press, Cambridge, 1991).
- Chang, T., Schiff, S. J., Sauer, T., Gossard, J.-P. & Burke, R. E. *Biophys. J.* **67**, 671–683 (1994).
- Schiff, S. J., Jerger, K., Chang, T., Sauer, T. & Aitken, P. G. *Biophys. J.* **67**, 684–691 (1994).
- Jefferys, J. G. R. *J. Physiol.* **319**, 143–152 (1991).
- Durand, D. *Brain Res.* **382**, 139–144 (1986).
- Kayali, H. & Durand, D. *Exp. Neurol.* **113**, 249–254 (1991).
- Ott, E. *Chaos in Dynamical Systems* (Cambridge Univ. Press, Cambridge, 1993).
- Auerbach, D., Grebogi, C., Ott, E. & Yorke, J. A. *Phys. Rev. Lett.* **69**, 3479–3482 (1992).

ACKNOWLEDGEMENTS. We thank P. G. Aitken for his invaluable teaching and advice and S. Mahan for research assistance. We gratefully acknowledge support from an NIMH grant (S.J.S.), the Children's Research Institute (S.J.S. and T.C.), an Office of Naval Research Young Investigator Award (W.L.D.), the Naval Surface Warfare Center's Independent Research Program (M.L.S.) and the Physics Division of the Office of Naval Research (M.L.S. and W.L.D.). The contribution of software from Manugistics Inc. is greatly appreciated.

Structure at 2.8 Å resolution of F₁-ATPase from bovine heart mitochondria

Jan Pieter Abrahams, Andrew G. W. Leslie, René Lutter* & John E. Walker†

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

* Present address: Academic Medical Centre, Department of Pulmonology, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

† To whom correspondence should be addressed

In the crystal structure of bovine mitochondrial F₁-ATPase determined at 2.8 Å resolution, the three catalytic β -subunits differ in conformation and in the bound nucleotide. The structure supports a catalytic mechanism in intact ATP synthase in which the three catalytic subunits are in different states of the catalytic cycle at any instant. Interconversion of the states may be achieved by rotation of the $\alpha_3\beta_3$ subassembly relative to an α -helical domain of the γ -subunit.

ADENOSINE triphosphate synthase (ATP synthase, F₁F₀-ATPase) is the central enzyme in energy conversion in mitochondria, chloroplasts and bacteria (reviewed in refs 1–6). It uses a protonmotive force, generated across the membrane by electron flow, to drive the synthesis of ATP from ADP and inorganic phosphate⁷. The multisubunit assembly consists of a globular domain, F₁, and an intrinsic membrane domain, F₀, linked by a slender stalk about 45 Å long. The F₁ domain is an approximate sphere 90–100 Å in diameter, and contains the catalytic binding sites for the substrates ADP and inorganic phosphate. Energy released by proton flux through F₀ is relayed to the catalytic sites in the F₁ domain, probably by conformational changes through the stalk. About three protons flow through the membrane per ATP synthesized.

Disruption of the stalk releases the water soluble enzyme, F₁-ATPase, a complex of five different proteins with the stoichiometry 3 α :3 β :1 γ :1 δ :1 ϵ . In bovine mitochondria they contain 510, 482, 272, 146 and 50 amino acids, respectively⁸, giving a relative molecular mass of 371,000. The sequences of the α - and β -subunits are homologous (20% identical), including the P-loop nucleotide-binding motif⁹. The catalytic sites are in the β -subunits, whereas the function of nucleotides in the α -subunits, which do not exchange during catalysis, is obscure. According to a binding change mechanism of ATP synthesis^{1,10}, the structures of the three catalytic sites are always different, but each passes through a cycle of 'open', 'loose' and 'tight' states. The mechanism suggests that F₁-ATPase is an inherently asymmetric assembly, as clearly indicated by subunit stoichiometry, electron

microscopy^{11,12} and low-resolution X-ray crystallography¹³. However, in one model proposed for rat F₁-ATPase based on analysis of a crystal form that implies three-fold symmetry, the three different α - and β -subunits are indistinguishable¹⁴.

To obtain a fuller understanding of the mechanism of F₁ and of the non-equivalence of the catalytic sites, we have determined the structure of F₁-ATPase from bovine heart mitochondria by X-ray diffraction at 2.8 Å resolution. Bovine F₁-ATPase is the largest asymmetric structure solved at atomic resolution so far. The α - and β -subunits, which have a similar fold, are arranged alternately like the segments of an orange around a central α -helical domain containing both the N- and C-terminal regions of the γ -subunit. As the three β -subunits contain different nucleotides and have different conformations, the structure as found in the crystal is compatible with one of the states to be expected in the cyclical binding change mechanism. It is likely that the central helical domain is part of a mechanism for transferring energy from the F₀ domain into the catalytic sites, and that in the process the 3 α -3 β assembly (referred to below as $\alpha_3\beta_3$) may rotate about the central α -helical domain of the γ -subunit.

Structure determination

Crystallization of bovine F₁ and data collection procedures have been described^{13,15}. The quality of data and initial phases determined by single isomorphous replacement and anomalous scattering are summarized in Table 1, and a representative sample of the electron density map is shown in Fig. 1. The model (Fig.

FIG. 1 Stereo illustration of part of the 3.2 Å-resolution electron density map used to build the initial model of bovine F₁-ATPase, with the refined model coordinates superimposed. The region shown corresponds to part of the β -sheet in the nucleotide-binding domain of subunit β_{DP} . The contour level is 1.2 times the r.m.s. density of the map.

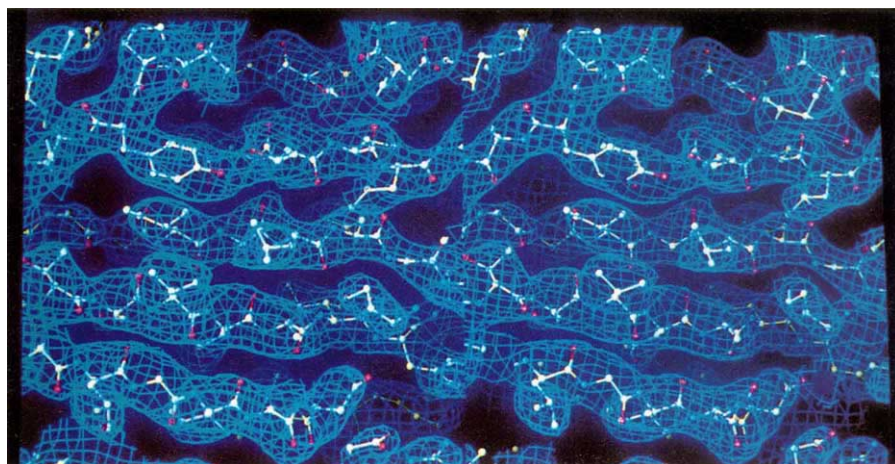


TABLE 1 Crystallographic data

	Data set 1		Data set 2		Data set 3	
	Native	0.2 mM methylmercury	Native	1 mM methylmercury	Native	0.2 mM methylmercury
Resolution	2.85 Å	3.2 Å	3.5 Å	3.5 Å	3.5 Å	3.5 Å
R_{merge}^*	0.098 (0.19)	0.077 (0.145)	0.059 (0.106)	0.067 (0.111)	0.114 (0.255)	0.086 (0.210)
Completeness	97.5% (92%)	96.3% (95.6%)	97.7% (87.7%)	97.4% (90.4%)	97.5% (99.8%)	93.5% (91.5%)
Multiplicity	5.4 (2.9)	3.7 (3.5)	3.2 (2.8)	7.5 (6.4)	8.9 (7.8)	2.4 (2.0)
No. of crystals	17	6	1	6	12	3
Rejected outliers†	0.37%	0.02%	0.56%	0.4%	0.3%	0.4%
Mean $\langle F/\sigma(F) \rangle$	30.9 (7.8)	28.8 (11.8)	46.9 (24.8)	53.8 (30.0)	32.6 (15.1)	19.2 (8.9)
$R_{\text{deriv}}^\ddagger$		0.114 (0.164)		0.120 (0.132)		0.170 (0.215)
No. of mercury sites		4		9		4
Total occupancy§		2.4		3.0		2.4
$R_{\text{cutoff}}^ $		0.60 (0.80)		0.56 (0.71)		0.72 (0.87)
Phasing power¶		1.37 (0.95)		2.0 (1.6)		1.08 (0.78)

Diffraction data. As native crystals were sometimes non-isomorphous, every time a derivative data set was collected, a corresponding native data set was collected too. Data sets 1 and 2 were collected on stations 9.6 and 9.5 respectively at the Synchrotron Radiation Source, Daresbury, UK. Data set 3 was collected using an Elliot GX11 generator with a 100 μ focal spot, collimated by a Supper double-mirror system. Data sets were recorded on MAR Research image-plate detectors, with a diameter of 300 mm for data sets 1 and 3, or 180 mm for data set 2. Data were processed with MOSFLM⁴³ and programs from the CCP4 suite⁴⁴. Values given in parenthesis are for the highest resolution bin. The R_{deriv} between native data sets 1 and 3 was 0.074, that between data sets 1 and 2 was 0.157, and that between data sets 2 and 3 was 0.162, indicating that data set 2 was not isomorphous with data sets 1 and 3, and that data sets 1 and 3 were not isomorphous to a lesser degree. Data from derivatized crystals were merged only if refinement of the heavy-atom parameters of the individual crystals indicated that this was allowed; otherwise they were treated separately. The phase-probability distributions of different data sets were combined, but only the structure-factor amplitudes of the high-resolution data set were used in the structure determination. **Solvent density modification.** Phases to 3.2 Å were improved by a new type of solvent density modification, to be described elsewhere. The solvent region (50% of the crystal volume) was determined from the local variation in electron density within a sphere of radius 3.7 Å. Features in the solvent region were inverted with respect to the mean solvent density. The phases of the resulting map were combined with the original phase probability distribution using σ_A weighting⁴⁵, and a new map was calculated. Iteration of this process decreased the mean phase error from 70 to 40°. Phase improvement was monitored by measuring the reduction in variation of the electron density within an unmodified spherical patch of solvent with a radius of 8 Å. **Model building and refinement.** The atomic model was built using program O (ref. 46). The few ambiguities in following the path of the polypeptide backbone of the α - and β -subunits were resolved by comparing the electron density of different copies of these subunits. Conventional refinement using XPLOR-3.1 (ref. 47) resulted in a model with an R -factor of 19.4% for 95% of the data between 6.0 and 2.86 Å. The free R -factor⁴⁸ for the remaining 5% of the data within this resolution range was 28.5%. The r.m.s. deviations from standard values of bond lengths and angles are 0.016 Å and 2.1°.

* $R_{\text{merge}} = \sum_i \sum_h |I(h) - \langle I(h) \rangle| / \sum_i \sum_h I(h)_i$, where $I(h)$ is the mean intensity after rejections. The contributing reflections were weighted by their standard deviations, which were determined by adjusting the measured standard deviation to reflect the observed differences between symmetry related reflections.

† Reflections with intensities differing more than 3.5 $\sigma(I)$ from the weighted mean were rejected.

‡ $R_{\text{deriv}} = \sum_i ||F_{\text{PH}}| - |F_P|| / \sum_i |F_P|$, where F_{PH} is the structure factor of the derivative and F_P is that of the native data.

§ The total refined occupancy of the sites, normalized to the site with the highest occupancy after setting this to 1.

|| $R_{\text{cutoff}} = \sum_i ||F_{\text{PH}} - F_P| - |F_{\text{H(calc)}}|| / \sum_i ||F_{\text{PH}} - F_P|$; F_P and F_{PH} are defined above, $F_{\text{H(calc)}}$ is the calculated heavy-atom structure factor. The summation is over centric reflections only.

¶ R.m.s. isomorphous difference divided by the r.m.s. residual lack of closure.

2) contains 2,983 amino acids in the following sequences: α_{DP} 19–510, α_{TP} 25–510, α_{E} 24–510, 9–474 in all three β -subunits, γ 1–45, γ 73–90 and γ 209–272 (see Fig. 2 legend for subunit nomenclature). Four AMP-PNP molecules and one ADP are also present (where AMP-PNP is 5'-adenylyl-imidodiphosphate). No water molecules were modelled into the structure. The δ - and ϵ -subunits and 145 residues of the γ -subunit were not located in the electron density map.

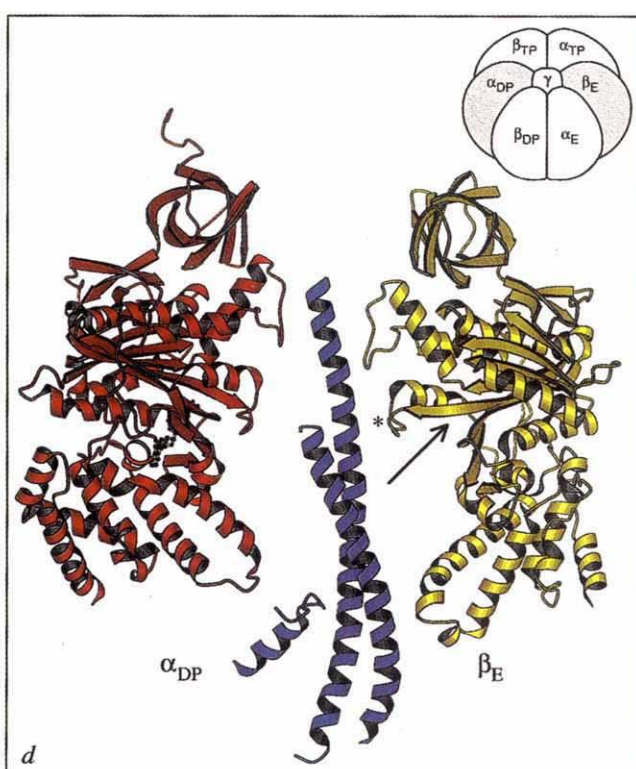
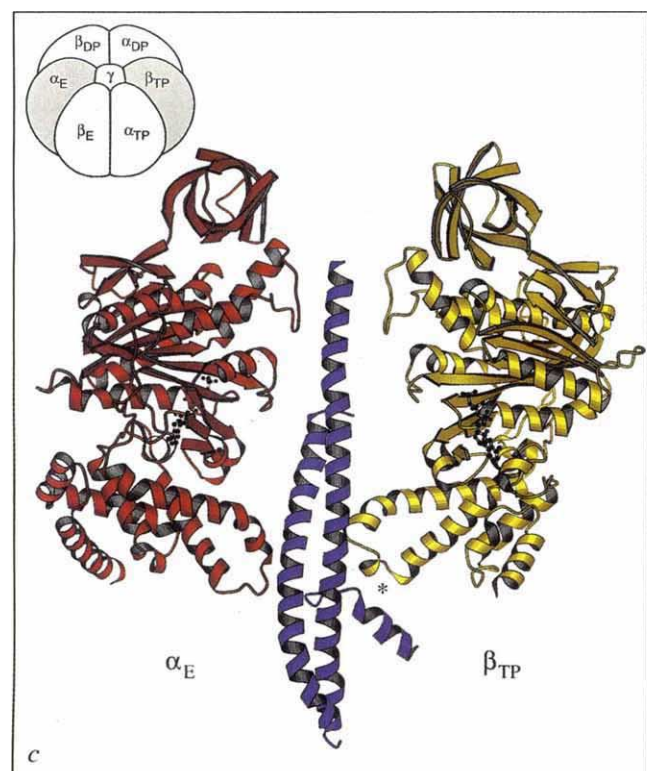
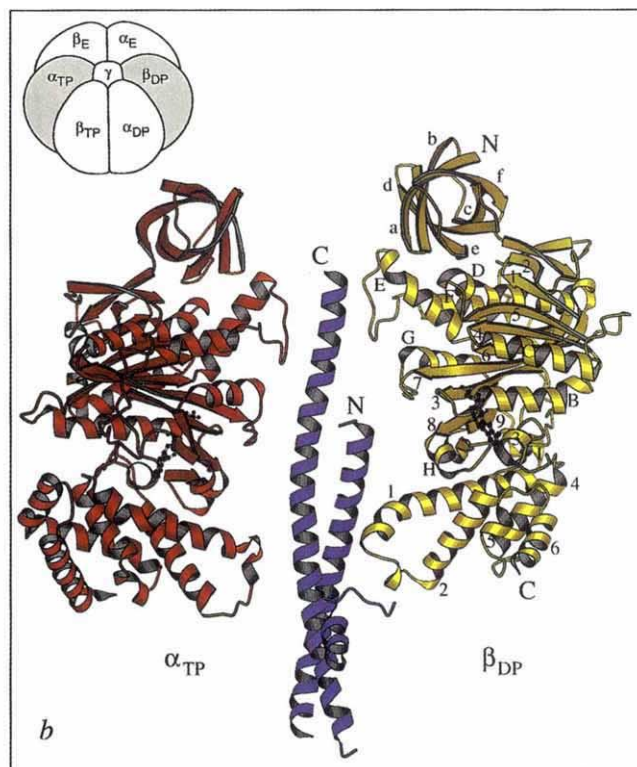
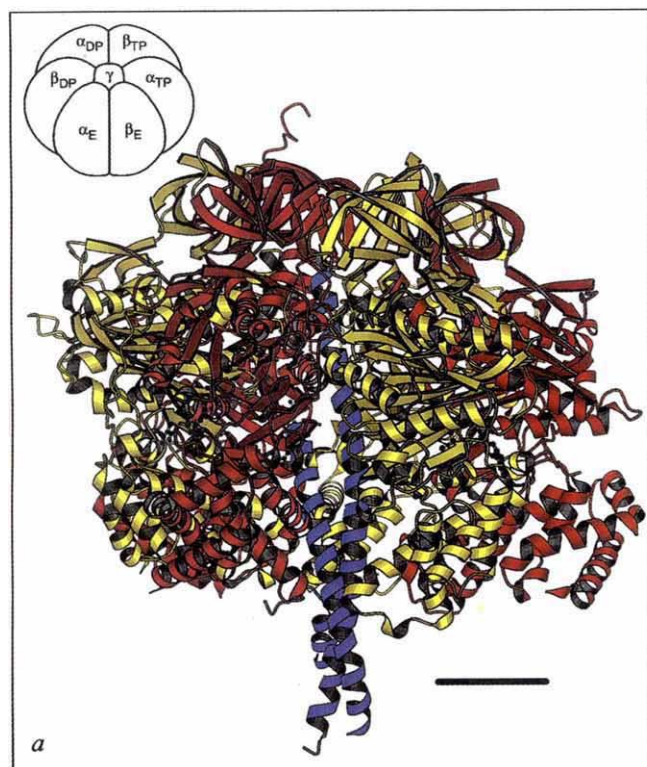
It was possible to solve the structure with data from a single heavy-atom derivative as a result of recent advances in detector technology. Apart from the new solvent density modification method (Table 1), standard techniques were used.

Architecture of F_1 -ATPase

F_1 is a flattened sphere 80 Å high and 100 Å across. The three α - and three β -subunits are arranged alternately like the segments of an orange around a central α -helix 90 Å long formed by the C-terminal amino acids γ 209–272 (Fig. 2). The C-terminal residue of this helix emerges into a dimple 15 Å deep in the surface at the top of the assembly. The lower half of the helix forms a left-handed antiparallel coiled coil with a second α -helix composed of amino acids γ 1–45. This helical structure protrudes about 30 Å from the bottom of the main body in a stem, almost certainly a vestige of the 45 Å stalk in ATP synthase. An increased local variation of the electron density suggests that the δ - and ϵ -subunits are probably in the stem region, but are not

well ordered. However, it cannot be excluded that the δ - and ϵ -subunits are sub-stoichiometric in the crystals. The γ -subunit passes through a large internal cavity connected to the outside by crevices between neighbouring α - and β -subunits. In the low-resolution structure, this cavity was interpreted to be a pit¹³. The C-terminal helix of the γ -subunit is twisted (Fig. 2). At the top

FIG. 2 The three-dimensional structure of bovine F_1 -ATPase. In this figure and Fig. 3, parts a–c were prepared with MOLSCRIPT⁴⁹. The α -, β - and γ -subunits are red, yellow and blue, respectively, and nucleotides are black, in a 'ball-and-stick' representation. The axis of pseudo-symmetry is vertical. AMP-PNP is bound to the three α -subunits, and to the β -subunit defined as β_{TP} . Subunit β_{DP} has bound ADP and subunit β_{E} has no associated nucleotide. The relationships of the various α - and β -subunits to each other is summarized in the top left or right corner (a–d). In b–d, the shaded part of the diagram indicates which subunits are depicted. Subunit α_{TP} contributes to the nucleotide-binding site of β_{TP} , and similarly for α_{DP} and α_{E} . Subunits α and γ are numbered from 1–510 and 1–272, respectively. By convention, the fifth amino acid (serine) in subunit β is residue 1 and the first four amino acids (AlaAlaGlnAla) are referred to as residues –1 to –4. The C-terminal amino acid is residue 478. a, A view of the entire F_1 particle in which subunits α_{E} and β_{E} point towards the viewer, revealing the antiparallel coiled coil of the N- and C-terminal helices of the γ -subunit through the open interface between them. The bar is 20 Å long. b, Subunits α_{TP} , γ and β_{DP} . The N and C termini of the β - and γ -subunits are shown. The



β -sheets in the N-terminal β -barrel are labelled a–f, consecutively. The order of helices (A–I) and the β -sheets (1–9) in the nucleotide-binding domain is as follows: 1, 2, A (not visible, hidden by helix 3 from the C-terminal domain), 3, B, 4, C, 5, D, 6, E, F, 7, G, 8, H, 9, I. The helices in the C-terminal domain are labelled 1–6, consecutively. The α -subunits have the same fold, with a minor deviation in the C-terminal domain, where the equivalent of helix 3 in the β -subunit is missing, and where two more helices are present after the equivalent of helix 6. *c*, Subunits α_E , γ and β_{TP} . The asterisk indicates a 'catch' interaction

of the loop containing the DELSEED sequence and the γ -subunit. *d*, Subunits α_{DP} , γ , and β_E . The arrow indicates the disruption of the β -sheet in the nucleotide-binding domain. β_E -Asp 316, β_E -Thr 318 and β_E -Asp 319 in a loop of the nucleotide-binding domain hydrogen-bonds with residues γ -Arg 254 and γ -Gln 255 from the C-terminal helix, 6 Å below the hydrophobic sleeve. The asterisk indicates the loop that makes a 'catch' with the C-terminal part of the γ -subunit. *e*, *f*, Stereo α -carbon backbone drawings of subunits α_{TP} and β_{DP} , respectively, in the same orientations as in the β -subunit in *b*.

of the assembly, its axis is coincident with the pseudo-threefold axis relating the α - and β -subunits, but at the lower end of the stem it is displaced from the axis by 7 Å. A third segment of the γ -subunit (γ 73–90) forms a small α -helix and loop in the region of the stem (Fig. 2).

The folds of the α - and β -subunits are almost identical, so that elements of three- and six-fold symmetry are present in the structure. Each subunit consists of an N-terminal six-stranded β -barrel (α 19–95, β 9–82), a central α – β domain containing the nucleotide-binding site (α 96–379, β 83–363), and a C-terminal bundle (α 380–510, β 364–474) of 7 and 6 helices, respectively,

in the α - and β -subunits (Fig. 2). In each subunit, the β -barrel and nucleotide-binding domains are separated, whereas the nucleotide-binding domain and the helical bundle interdigitate to a minor extent. The extent of contact between neighbouring α - and β -subunits varies (Fig. 3). The β -barrel domains crown the particle with a continuous 24-stranded β -sheet (Fig. 3a), whereas the shortest distance between side chains of some of the helical bundles (Fig. 3c) is 10 Å. Between 18 and 24 amino acids at the N termini of α -subunits and amino acids β (–4 to 8) are disordered. In F_1 -ATPase and in ATP synthase, these regions are protease-sensitive^{8,16}.

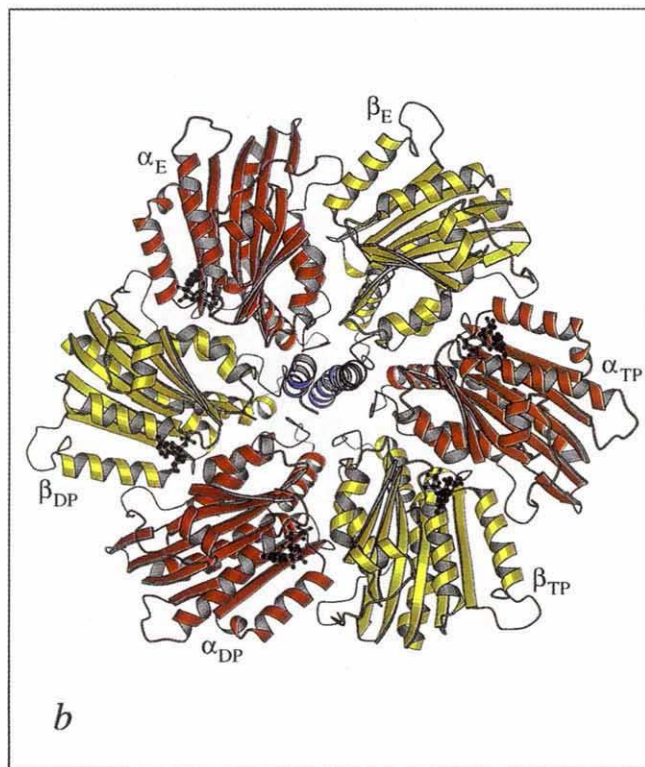
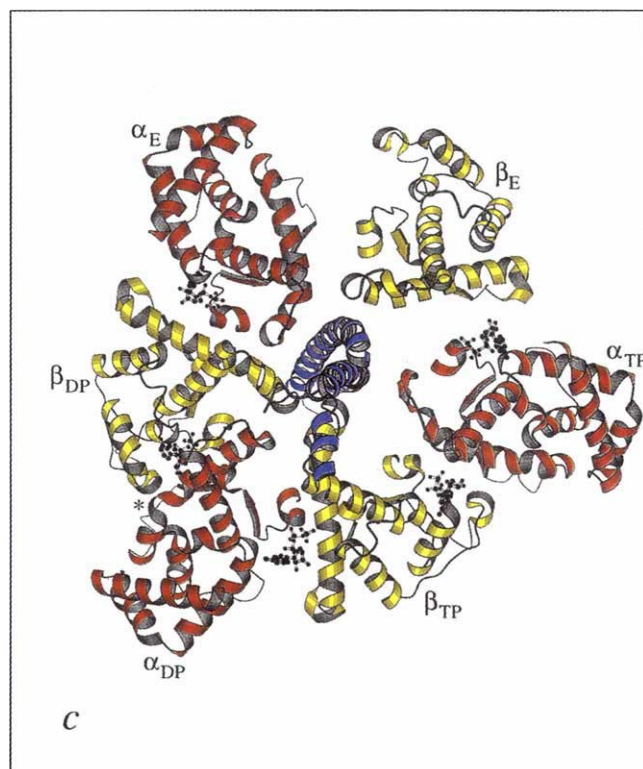
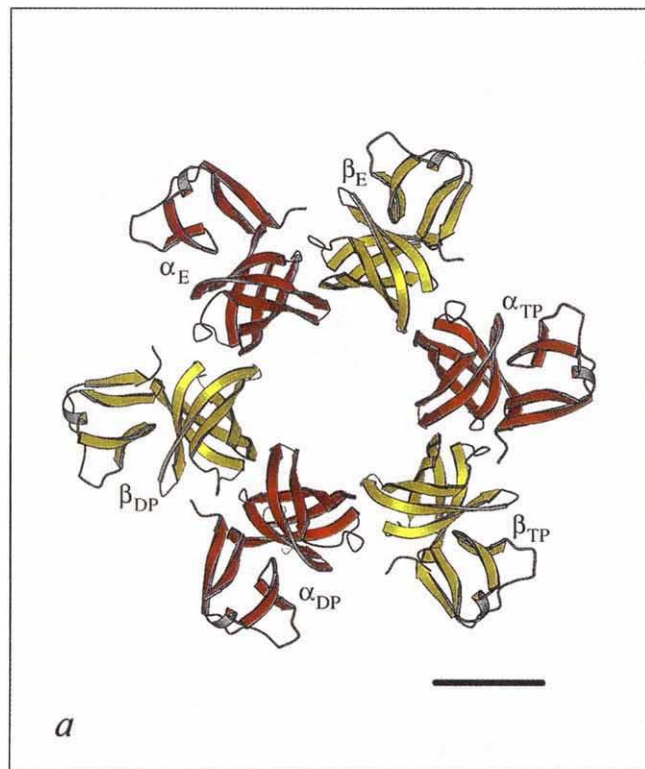
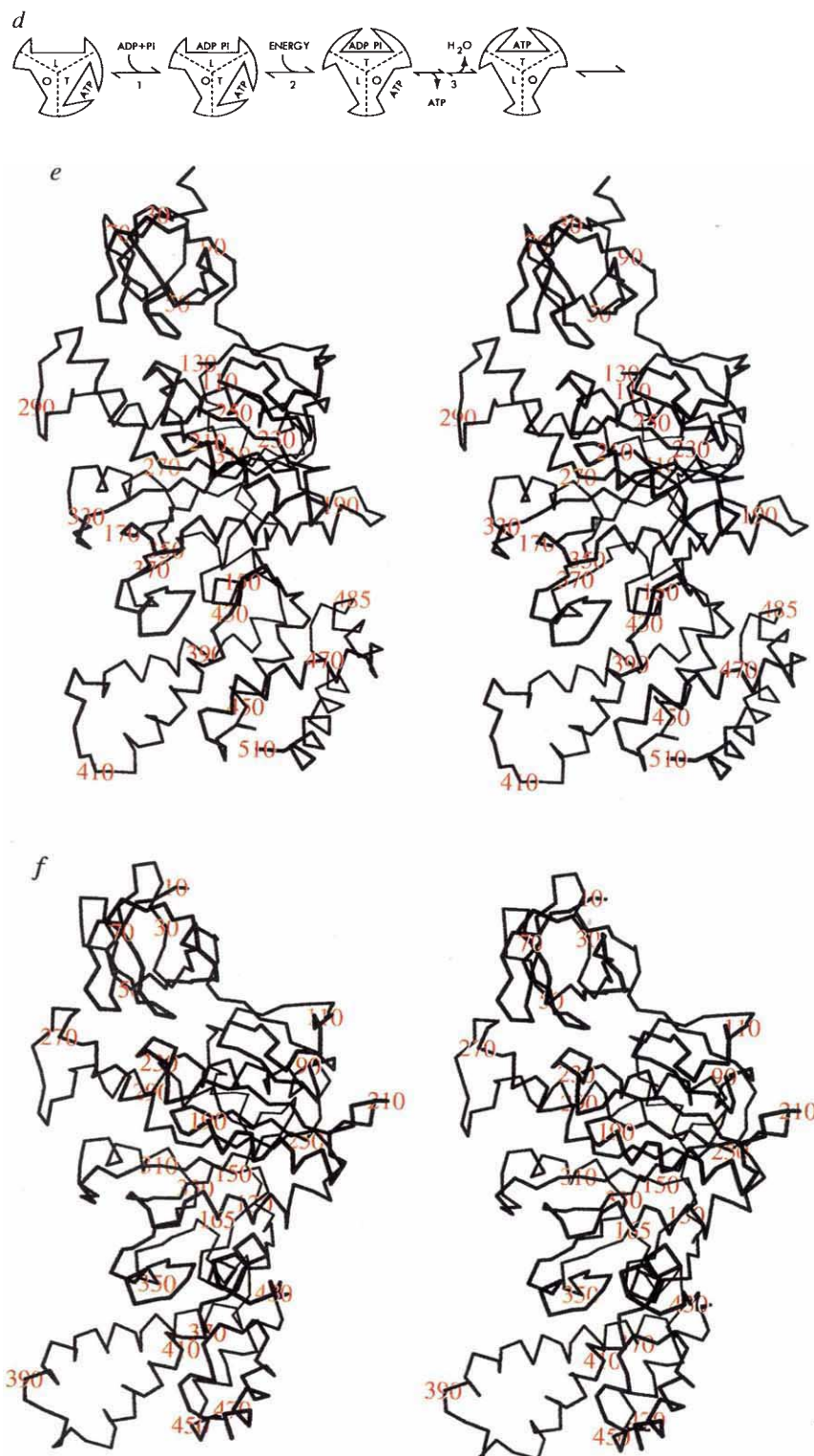


FIG. 3 Asymmetrical features in bovine F_1 -ATPase and the binding change mechanism; a–c are viewed from the membrane side, the three views being drawn to the same scale and in the same orientation (rotated by 90° relative to Fig. 2a). a, The N-terminal domains and part of the nucleotide-binding domains of α - and β -subunits. The N-terminal domains make a 24-stranded β -barrel with the six α -strands hydrogen-bonded to β -strands of the neighbouring subunit (Fig. 2b). Scale bar, 20 Å. b, The nucleotide-binding domains. The β -sheet of subunit β_E is disrupted close to the γ -subunit. c, The C-terminal domains. The deviations from perfect symmetry (most apparent in this view) result in the characteristic features of the individual nucleotide-binding sites. d, An energy-dependent binding change mechanism of ATP synthesis. The catalytic sites in β -subunits interact and interconvert between three forms: O, open site with very low affinity for ligands and catalytically inactive; L, loose site, loosely binding ligands and catalytically inactive; T, tight site, tightly binding ligands and catalytically active. The proton-induced conformational changes convert a T-site with bound ATP to an O-site, releasing the bound nucleotide. Concomitantly, an L-site with loosely bound ADP and phosphate is converted to a T-site, where the substrates are tightly bound and ATP forms. Fresh substrates bind to an O-site, converting it to an L-site, and so on. The energy-requiring steps are substrate binding and product release. One third of the catalytic cycle is shown. (Reproduced with permission from ref. 1).

The nucleotide-binding domain contains a nine-stranded β -sheet with nine associated α -helices. In both α - and β -subunits, β -strands 3–9 and helices A–G have the same topology as β strands 1–7 and associated helices in RecA¹⁷. Within this region, between 113 and 145 α -carbons superimpose on RecA with an rms separation of 1.9–2.0 Å (depending on the subunit). The nucleotide pyrophosphate group binds in the same way in both structures, but the orientation of the adenosine moiety is very different. In adenylate kinase¹⁸, often used as a model for F_1 -ATPase, the corresponding β -sheet has five strands with different connectivity.

Nucleotide-binding sites

The nucleotide-binding sites are at the interfaces between α - and β -subunits. The catalytic sites are predominantly in β -subunits, with some contributions from side chains in the α -subunits, and conversely with the non-catalytic sites. Most notably, β -Tyr 368, a site of reaction of both photoactivated 2-azido-ADP¹⁹ and of the nucleotide analogue 5'-p-fluorosulphonylbenzoyl-adenosine²⁰, protrudes into the α -subunit nucleotide site (Fig. 4). The average distance between adjacent catalytic and non-catalytic sites is 27 Å (between β -phosphates). In both α - and β -subunits, one face of the ribose and phosphate moieties is



accessible to external solvent via a conical tunnel. In the crystals, the subunit denoted β_{TP} and each of the three α -subunits bind one molecule of the nucleotide triphosphate analogue AMP-PNP, whereas the diphosphate ADP is bound to β_{DP} . In subunit β_E , the nucleotide-binding site is empty and distorted, despite the presence of 250 μ M AMP-PNP and 5 μ M ADP in the solvent¹⁵. Inorganic phosphate is not present in the ADP site, and so the structure is probably that of the ADP-inhibited state¹⁰. The three α -subunits are defined in relation to β_{TP} , β_{DP} and β_E . Subunit α_{TP} forms the interface with β_{TP} that contains the β_{TP} nucleotide site, and similarly for α_{DP} and α_E (Fig. 2).

The amino acids involved in nucleotide binding are summarized in Fig. 4. Those that define the adenine- and ribose-binding pocket in the β -subunits include β -Tyr 345, which forms a crosslink with photoactivated 2-azido-ATP²¹. In contrast, β -Tyr 311, which reacts with photoactivated 8-azido-ATP^{22,23}, is 5.5 Å from the terminal phosphate at the opposite end of the binding site to the adenine pocket. The reaction appears to require non-productive binding of the nucleotide. The most important phosphate-binding amino acids are in the P-loop, which can be seen between strand 3 and helix B in subunit β_{DP} in Fig. 2b. This loop contains the conserved nucleotide-binding motif, GXXXXGKT/S, first recognized in the sequences of the α - and β -subunits of F_1 -ATPase⁹. The motif has been used to predict the nucleotide-binding sites of many other proteins. The crystal structures of p21^{ras} (ref. 24), adenylate kinase²⁵, RecA²⁶, elongation factor Tu^{27,28} and transducin- α ²⁹, for example, show that the motif binds the di- or triphosphate moiety, and this is also true in the α - and β -subunits of F_1 -ATPase (Fig. 4).

The crystal structure provides new details of the chemical mechanism of ATP synthesis. In β_{TP} , at a distance of 4.4 Å from the γ -phosphate of AMP-PNP, there is clear density for a bound water molecule, hydrogen-bonded to the carboxylate of β -Glu 188. This residue is appropriately positioned to activate the water molecule, promoting an in-line nucleophilic attack on the terminal phosphate. The guanidinium of α -Arg 373 could help to stabilize the negative charge that develops on the terminal phosphate in a pentacoordinate transition state. A similar arrangement of a glutamic acid and an arginine side chain is found in the active site of transducin- α complexed with GTP- γ S²⁹. The structure also indicates why the α -subunits

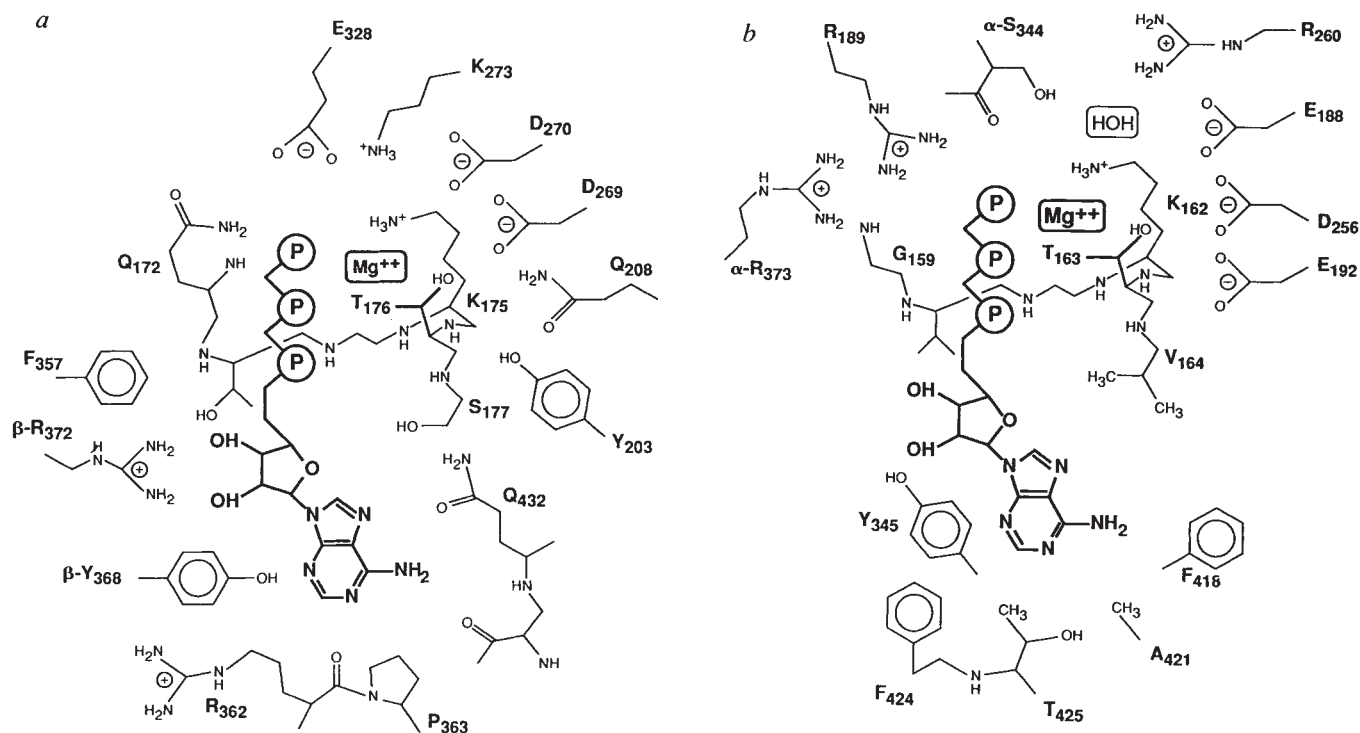


FIG. 4 The nucleotide-binding sites in the α - and β -subunits of bovine F_1 -ATPase. *a*, Subunit α_E . All of the amino acids forming the pocket are in the α -subunit, except for the neighbouring β_{DP} -Arg 372 and β_{DP} -Tyr 368. Residues $\alpha_{172-177}$ are part of the P-loop. The nucleotide-binding sites of α_{DP} and α_{TP} are very similar, except that β -Arg 372 does not contribute to either binding site, and β -Tyr 368 only contributes to nucleotide binding in α_{DP} . *b*, Subunit β_{TP} . Except for α_{TP} -Arg 373 and the main-chain carbonyl and the side chain of α_{TP} -Ser 344, all of the amino acids are in β_{TP} . Residues 159–164 are part of the P-loop. In β_{DP} , there is density for an ordered water molecule replacing the

terminal phosphate, and Arg 373 moves closer to the diphosphate moiety. Except for β -Arg 372 in the α -subunit nucleotide-binding site, all residues forming the pockets are widely conserved. The ligands to magnesium identified at this stage are as follows: in the α -subunits, O_{γ} -Thr 176, two oxygens from the terminal phosphate and one from the β -phosphate; in β_{TP} , O_{γ} -Thr 163 and one oxygen atom from each of the β - and terminal phosphates; in β_{DP} , O_{γ} -Thr 163 and one oxygen from the β -phosphate. The remaining ligands are presumably water molecules.

are non-catalytic. In the α -subunits, there is no spatial equivalent of the carboxylate of β -Glu 188, which is replaced in α -subunits by Gln 208, with the side chain pointing away from the terminal phosphate.

The binding of adenine by the α -subunit involves several hydrogen bonds, whereas in the β -subunit the adenine is in contact with a hydrophobic interface. Together with the presence of β -Tyr 368 close to the 2-position of the adenine ring in the α -subunit, this may explain the specificity of the α -subunit nucleotide sites for adenine nucleotides, and the ability of the β -subunit nucleotide sites to accommodate GTP and ITP as well as ATP^{6,30}. The roles of the non-exchangeable nucleotides in the α -subunits remain obscure.

The site of reaction of the inhibitor dicyclohexylcarbodiimide (β -Glu 199; ref. 31) points into the conical tunnel leading to the nucleotide sites. The presence of the two bulky cyclohexyl moieties in the conical tunnel could impede nucleotide binding and the conversion of the catalytic sites between conformational states.

Asymmetrical features

With the exception of the nucleotide-binding domain of β_E , the domains of the three α - and three β -subunits superimpose with a root-mean-squared separation of less than 1 Å between C α atoms, but because of differences in the relative orientations of the domains and the interactions with the unique γ -subunit, the overall structure of F_1 is strikingly asymmetric (Fig. 3).

In the β_E nucleotide-binding domain, the interaction between strands 3 and 7 is disrupted (above and below the arrow in Fig. 2d). The lower part of this domain and the α -helical domain are rotated by 30° with respect to their positions in β_{TP} and

β_{DP} , resulting in movements by up to 20 Å. At the C terminus of helix B following the P-loop (below the arrowhead in Fig. 2d), the disruption of the β -sheet is accompanied by relaxation of a few residues into an α -helical conformation. A number of other hydrogen-bonding rearrangements occur throughout the rest of β_E . Also, β_E -Asp 316, β_E -Thr 318, and β_E -Asp 319 in the loop following strand 7 (indicated by the asterisk in Fig. 2d) hydrogen-bond with γ -Arg 254 and γ -Gln 255 in the C-terminal helix. Close by, α_E -Asp 333 makes a salt bridge with γ -Arg 252. These H-bonds form a 'catch', and are the major specific interactions between the two long helices of the γ -subunit and the subassembly $\alpha_3\beta_3$.

A second 'catch' is found in the C-terminal domain of subunit β_{TP} , but not in β_{DP} or β_E . Additional hydrogen bonds link γ -Lys 87, γ -Lys 90 and the backbone amide of γ -Ala 80 (all in the short helix, which is inclined at about 45° to the axis of pseudosymmetry) with β_{TP} -Asp 394 and β_{TP} -Glu 398 in a loop between helices 1 and 2. This sequence (DELSEED, β 394–400) is part of the binding site of amphipathic cationic inhibitors^{32,33} and possibly of the ATPase inhibitor protein³⁴ in one or other of the β -subunits.

Asymmetry is also evident in α -subunits. The non-catalytic nucleotide-binding sites in α_{DP} and α_E are very similar, but in α_{TP} the site is more open because the lower half of β_E (which contributes to this binding site) is hinged out. Therefore, β_E -Tyr 368 no longer interacts with the N3 atom of the adenosine. It is likely that α_{TP} binds the exchangeable non-catalytic nucleotide observed in nucleotide-binding studies¹⁰.

Asymmetry is also manifest in the coiled coil of the γ -subunit. Its C terminus (γ 265–272) passes through a hydrophobic sleeve

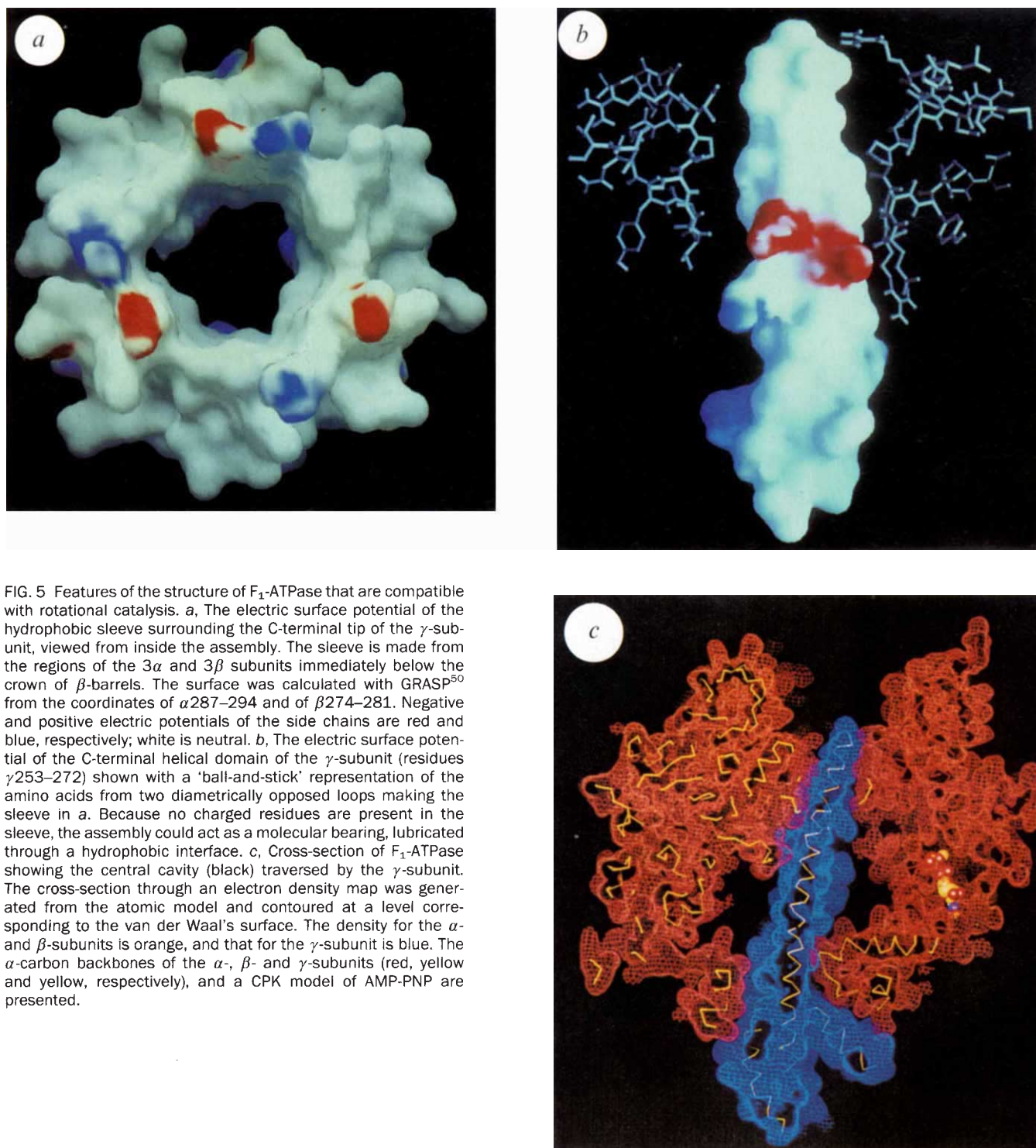


FIG. 5 Features of the structure of F_1 -ATPase that are compatible with rotational catalysis. *a*, The electric surface potential of the hydrophobic sleeve surrounding the C-terminal tip of the γ -subunit, viewed from inside the assembly. The sleeve is made from the regions of the 3α and 3β subunits immediately below the crown of β -barrels. The surface was calculated with GRASP⁵⁰ from the coordinates of $\alpha 287$ – 294 and of $\beta 274$ – 281 . Negative and positive electric potentials of the side chains are red and blue, respectively; white is neutral. *b*, The electric surface potential of the C-terminal helical domain of the γ -subunit (residues $\gamma 253$ – 272) shown with a 'ball-and-stick' representation of the amino acids from two diametrically opposed loops making the sleeve in *a*. Because no charged residues are present in the sleeve, the assembly could act as a molecular bearing, lubricated through a hydrophobic interface. *c*, Cross-section of F_1 -ATPase showing the central cavity (black) traversed by the γ -subunit. The cross-section through an electron density map was generated from the atomic model and contoured at a level corresponding to the van der Waal's surface. The density for the α - and β -subunits is orange, and that for the γ -subunit is blue. The α -carbon backbones of the α -, β - and γ -subunits (red, yellow and yellow, respectively), and a CPK model of AMP-PNP are presented.

formed by six proline-rich loops ($\alpha 287$ – 294 , $\beta 274$ – 281 ; following helix E in β_{DP} in Fig. 2*b*, and the related loop in α_{TP}). The remaining interactions with the γ -subunit take place through hydrophobic loops, some of them poorly ordered, in the C-terminal domains of the α - and β -subunits, and helix 1 in β_{DP} . In the crystal structure, the γ -subunit seems to prevent β_E from adopting a nucleotide-binding conformation by obstructing the rotation of its lower half towards the central axis. In consequence, the lower halves of β_E and α_{TP} do not interact, and the middle and C-terminal domains of α_{TP} have moved away from the central axis (Fig. 3*c*). Therefore, each β -subunit adopts a different orientation towards the α -subunit that contributes to its active site, and the active site interfaces of β_{DP} , β_{TP} and β_E

appear from a structural point of view to be tightly closed, more open, and fully open, respectively. This suggests that β_{DP} contains the non-exchangeable catalytic site¹⁰. The closure of the active site interface is accompanied by a distortion of helix 1 in the C-terminal domain of the α -subunits at Gly 388 (Fig. 3*c*). The preference of β_{TP} and β_{DP} for AMP-PNP and ADP, respectively, is probably caused by small conformational differences near to their P-loops, the most significant being the shift of α_{DP} -Arg 373 by 1.5 Å towards the β -phosphate moiety in β_{DP} .

Rotational catalysis

The asymmetry required by the binding change mechanism of ATP synthesis (Fig. 3*d*), particularly in the region of the active

sites, is evident in the crystal structure. The three different conformations of the subunits β_E , β_{TP} and β_{DP} could be identified on the structural grounds described above as representing the open, loose and tight states, respectively. This asymmetry is associated with the asymmetric positioning of the γ -subunit relative to the $\alpha_3\beta_3$ subassembly.

The nucleotides that are found in the catalytic sites in the crystals appear at first sight to be inconsistent with this proposal, as the tight site (β_{DP}) should be occupied by ATP, and the loose site (β_{TP}) by ADP or ADP plus phosphate. However, if, as we propose, the structure in the crystal is that of the ADP-inhibited form of the enzyme, ADP (but not P_i) would be bound at the tight catalytic site. The presence of AMP-PNP in the loose site may be due to its 50-fold higher concentration over ADP, and to the absence of P_i from the crystallization medium. Alternatively, it is possible that the sequence of events during catalysis differs from that shown in Fig. 1, and that the 'tight' site of newly synthesized ATP first becomes 'loose' and then 'open'. Crystals of F_1 grown in the presence of P_i and alternative nucleotides (ATP, ATP- γ S) could help to clarify this ambiguity.

The transition between the different catalytic states implied by the cycle of binding changes could be achieved by rotation of the catalytic subunits relative to the rest of the ATP synthase. This hypothesis cannot be tested by a single 'snapshot' of F_1 -ATPase provided by the crystal structure, but several features of the structure are compatible with rotation of the $\alpha_3\beta_3$ subassembly relative to the γ -subunit (and presumably the remainder of the intact synthase). First, the sleeve surrounding the C terminus of the γ -subunit looks like a molecular bearing, lubricated by a hydrophobic interface (Fig. 5a and b). The crown of β -barrels above it, which form a continuous β -sheet linking all six subunits, holds the structure together, and the top of the nucleotide domain also contributes to its stability. Second, the central cavity would permit passage of the long N-terminal helix of the γ -subunit within the core of F_1 during rotation, with minimal contact between the γ -subunit and the rest of the interior of the particle (Fig. 5c). Third, the differences between the three catalytic sites are correlated with the asymmetric position of the γ -subunit. If there is to be an interconversion of the

three sites, as the binding-change mechanism suggests, the most obvious way to achieve this is by rotation of $\alpha_3\beta_3$ relative to the γ -subunit. The nature of the interactions between the γ -subunit and $\alpha_3\beta_3$ suggest that the energy barrier opposing rotation is quite small. It could be overcome with energy from conformational changes induced by proton transport and relayed to the active sites via subunit γ and other stalk components. The V_{max} of the bovine ATP synthase is about 400 per second, and twice that value in the chloroplast enzyme¹⁰. Therefore, the rate of rotation in ATP synthases would be about 130–270 Hz.

Rotation of F_1 in ATP synthase and of the hydrophobic subunits in F_0 have been considered in the past^{10,35–37}. Attempts to demonstrate rotation have given negative results¹⁰, but the possibility of rotation should be re-examined in the light of the structure. In the F_1 -ATPase from *Escherichia coli*, nucleotide-dependent movements of the minor subunits (γ , δ and ϵ) in the central space between the hexameric arrangement of α - and β -subunits have been observed by cryoelectron microscopy¹² and by changes in the properties of fluorescent labels³⁸. In the chloroplast enzyme, there is also evidence of movement of a central mass³⁹. The proposed rotation within ATP synthase may be analogous to the rotation of the bacterial flagellum, which is also driven by the transmembrane protonmotive force⁴⁰. A homologue of the β -subunit of ATPase is associated with the flagella complex, and may be involved in its assembly⁴¹. Assessment of the validity of the analogy between ATPase and flagellae requires appropriate biophysical examination of the ATP synthase based on the structure of F_1 , and further detailed structural investigation of both ATP synthase and the flagellar motor. In the ATP synthase, it will be necessary to know more about the coupling mechanism (of which the γ -subunit and probably δ and ϵ are part) linking the catalytic sites with proton transport in the F_0 domain. The characterization of the proteins in the bovine stalk^{16,42} and the *in vitro* assembly of an F_1 /stalk complex⁴² are necessary steps towards that goal. Little is known about the structure of the intrinsic membrane domain, F_0 , and determining its high-resolution structure and understanding the structural basis of the coupling of proton transport to ATP synthesis are still challenges for the future. □

Received 21 June; accepted 19 July 1994.

1. Cross, R. L. *A. Rev. Biochem.* **50**, 681–714 (1981).
2. Senior, A. E. *Physiol. Rev.* **68**, 177–231 (1988).
3. Fillingame, R. H. *The Bacteria* **12**, 345–391 (1990).
4. Walker, J. E., Fearnley, I. M., Lutter, R., Todd, R. J. & Runswick, M. J. *Phil. Trans. R. Soc.* **326**, 367–378 (1990).
5. Peneffsky, H. S. & Cross, R. L. *Adv. Enzym.* **64**, 173–214 (1991).
6. Issartel, J. P., Dupuis, A., Garin, J., Lunardi, J., Michel, L. & Vignais, P. V. *Experientia* **48**, 351–362 (1992).
7. Mitchell, P. *Biol. Rev.* **41**, 455–502 (1966).
8. Walker, J. E. *et al. J. molec. Biol.* **184**, 677–701 (1985).
9. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. *EMBO J.* **1**, 945–951 (1982).
10. Boyer, P. D. *Biochim. biophys. Acta* **1140**, 215–250 (1993).
11. Boekema, E. J., Berden, J. A. & van Heel, M. G. *Biochim. biophys. Acta* **851**, 353–360 (1986).
12. Wilkens, S. & Capaldi, R. A. *Biol. Chem. Hoppe-Seyler* **375**, 43–51 (1994).
13. Abrahams, J. P. *et al. EMBO J.* **12**, 1775–1780 (1993).
14. Bianchet, M., Ysern, X., Hüllihen, J., Pedersen, P. L. & Amzel, M. J. *biol. Chem.* **266**, 21197–21201 (1991).
15. Lutter, R. *et al. J. molec. Biol.* **229**, 787–790 (1993).
16. Collinson, I. R., Fearnley, I. M., Runswick, M. J., Skehel, J. M. & Walker, J. E. *Biochem. J.* (in press).
17. Story, R. M., Weber, I. T. & Steitz, T. A. *Nature* **355**, 318–325 (1992).
18. Pai, E. F., Sachsenheimer, W., Schirmer, R. H. & Schulz, G. E. *J. molec. Biol.* **114**, 37–45 (1977).
19. Cross, R. L. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 5715–5719 (1987).
20. Esch, F. S. & Allison, W. S. *J. biol. Chem.* **253**, 6100–6106 (1978).
21. Garin, J., Boulay, F., Issartel, J. P., Lunardi, J. & Vignais, P. V. *Biochemistry* **25**, 4431–4437 (1986).
22. Hollemans, M., Runswick, M. J., Fearnley, I. M. & Walker, J. E. *J. biol. Chem.* **258**, 9307–9313 (1983).
23. Garin, J., Vinçon, M., Gagnon, J. & Vignais, P. *Biochemistry* **33**, 3772–3777 (1994).
24. Milburn, M. V. *et al. Science* **247**, 939–945 (1990).
25. Mueller, C. W. & Schultz, G. E. *J. molec. Biol.* **224**, 159–177 (1992).
26. Story, R. M. & Steitz, T. A. *Nature* **355**, 374–376 (1992).
27. Berchtold, H. *et al. Nature* **365**, 126–132 (1993).

28. Kjeldgaard, M., Nissen, P. O., Thirup, S. & Nyborg, J. *Structure* **1**, 35–50 (1993).
29. Noel, J. P., Hamm, H. E. & Sigler, P. B. *Nature* **366**, 654–663 (1993).
30. Perlín, D. S., Latchney, L. R., Wise, J. G. & Senior, A. E. *Biochemistry* **23**, 4998–5003 (1984).
31. Esch, F. S., Böhlen, P., Otsuka, A. S., Yoshida, M. & Allison, W. S. *J. biol. Chem.* **256**, 9084–9089 (1981).
32. Bullough, D. A., Ceccarelli, E. A., Roise, D. & Allison, W. S. *Biochim. biophys. Acta* **975**, 377–383 (1989).
33. Zhuo, S., Paik, S. R., Register, J. A. & Allison, W. S. *Biochemistry* **32**, 2219–2227 (1993).
34. Jackson, P. J. & Harris, D. A. *Biochem. J.* **235**, 577–583 (1986).
35. Cox, G. B., Jans, D. A., Fimmel, A. L., Gibson, F. & Hatch, L. *Biochim. biophys. Acta* **768**, 201–208 (1984).
36. Cox, G. B., Fimmel, A. L., Gibson, F. & Hatch, L. *Biochim. Biophys. Acta* **849**, 62–69 (1986).
37. Mitchell, P. *FEBS Lett.* **182**, 1–7 (1985).
38. Capaldi, R., Aggeler, R., Turina, P. & Wilkens, S. *Trends biochem. Sci.* **19**, 284–289 (1994).
39. Boekema, E. J. & Böttcher, B. *Biochim. biophys. Acta* **1098**, 131–143 (1992).
40. Macnab, R. M. *A. Rev. Genet.* **26**, 131–158 (1992).
41. Dreyfuss, G., Williams, A. W., Kawagishi, I. & Macnab, R. M. *J. Bact.* **175**, 3131–3138 (1993).
42. Collinson, I. R. *et al. J. molec. Biol.* (in press).
43. Leslie, A. G. W. in *Joint CCP4 and ESF-EACMB Newsletter on Protein Crystallography* No. 26 (Daresbury Laboratory, Warrington, UK, 1992).
44. CCP4: A Suite of Programs for Protein Crystallography (SERC Collaborative Computing Project No. 4, Daresbury Laboratory, Warrington, UK, 1979).
45. Read, R. J. *Acta crystallogr. A* **42**, 140–149 (1986).
46. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. *Acta crystallogr. A* **47**, 110–119 (1991).
47. Brünger, A. T. *XPLOR Version 3.1 Manual* (Yale Univ., USA, 1993).
48. Brünger, A. T. *Nature* **335**, 472–475 (1992).
49. Kraulis, P. J. *J. appl. Crystallogr.* **24**, 946–950 (1991).
50. Nicolls, A., Sharp, K. A. & Honig, B. *Proteins* **11**, 281–296 (1991).

ACKNOWLEDGEMENTS. We thank R. Henderson, A. Klug and M. F. Perutz for discussions, the staff of the Synchrotron Radiation facility at Daresbury for their help, and M. J. van Raaij and R. Noel for assistance. J.P.A. was supported during part of this work by an EMBO Fellowship. After completion of the current refinement, the coordinates will be deposited in the Brookhaven Protein Databank where they will be held for one year before release. We dedicate this paper to M. F. Perutz on the occasion of his 80th birthday.

Evidence against dissipationless dark matter from observations of galaxy haloes

Ben Moore*

Department of Astronomy, University of California, Berkeley, California 94720, USA

THERE are two different types of missing (dark) matter: the unseen matter needed to explain the high rotation velocities of atomic hydrogen in the outer parts of spiral galaxies^{1,2}, and the much larger amount of (non-baryonic) matter needed to prevent the universe from expanding forever¹ (producing either a 'flat' or a 'closed' Universe)³. Several models have been proposed to provide the dark matter required within galaxy haloes for a flat universe, of which cold dark matter (CDM) has proved the most successful at reproducing the observed large-scale structure of the Universe⁴⁻⁶. CDM belongs to a class of non-relativistic particles that interact primarily through gravity, and are named dissipationless because they cannot dissipate energy (baryonic particles can lose energy by emitting electromagnetic radiation). Here I show that the modelled small-scale properties of CDM⁷⁻⁹ are fundamentally incompatible with recent observations¹⁰⁻¹³ of dwarf galaxies, which are thought to be completely dominated by dark matter on scales larger than a kiloparsec. Thus, the hypothesis that dark matter is predominantly cold seems hard to sustain.

Inflationary and nucleosynthesis arguments have played a major role in advocating exotic particles as dark matter³; the former because inflation generally requires a flat Universe, and the latter because all the dark matter cannot be baryonic. These particles are called exotic because they have yet to be detected. They interact with baryons only via the gravitational force, and are therefore termed collisionless, as short-range particle-particle forces are not important. Depending on their present-day velocities, particles can be classed as either cold (for example, axions) or hot (for example, light neutrinos¹⁴). A universe dominated by hot dark matter (HDM) has difficulty in forming small-scale structure^{15,16} unless it is combined with a mechanism to seed galaxy formation, such as cosmic strings¹⁷. CDM-dominated universes have been extensively investigated numerically, and have proved quite successful at reproducing the observed clustering pattern on scales from galaxies to clusters of galaxies⁴⁻⁶. Structure formation in these models proceeds in a bottom-up fashion, with small over-densities in the mass distribution collapsing first and subsequently merging hierarchically to form larger objects. Very simply put, dwarf galaxies form first, and then merge to form larger galaxies. Galaxies form from gas that has cooled within the deep potential wells provided by the dark-matter haloes, plus the accretion of gas-rich subclumps which are part of the merging hierarchy. The dominant force which drives structure formation is gravity, and numerical simulations have proved a useful tool with which to study the non-linear growth of structure formation over a wide range of dynamical scales.

One of the first major successes apparent from the numerical simulations of cosmological models dominated by collisionless dark matter, was the formation around galaxy-sized structures of dark haloes whose density profiles fall approximately as r_g^{-2} , where r_g is the distance from the centre of the halo^{18,19}. This implied that the stars or gas would have a constant circular velocity at any position within the halo, $v_c(r_g) = (Gm_g/r_g)^{1/2}$, where m_g is the mass within r_g , in agreement with the observed rotation of gas in large spiral galaxies. The structure of dark haloes has recently been re-examined at a resolution 50 times

higher than the original simulations^{8,9}, and confirms the initial numerical investigations. Moreover, the density profiles increase to distances less than the resolution limit of ~ 1 kpc, a result which complements analytical studies of the formation of isolated spherical haloes⁷.

The density profiles of the simulated CDM haloes can be fitted relatively well by the empirical Hernquist profile⁸, which on scales of 10–50 kpc gives a density profile falling approximately as r_g^{-2} , and on smaller scales as r_g^{-1} . The resulting rotation curve is described by $v_c(r_g) = (Gm_s r_g)^{1/2} / (r_g + r_s)$, where the effective mass m_s can be taken as the mass within the virial radius, and r_s is a scale representing the peak height of the initial density fluctuation⁸. Dubinski and Carlberg⁸ simulated the formation of a large halo with peak circular velocity $v_{c,max} \approx 280$ km s⁻¹, and obtained values of $m_s = 2.0 \times 10^{12} M_\odot$ and $r_s = 26$ kpc. Warren *et al.*⁹ present the rotation curve of a halo from one of their simulations, which has a peak circular velocity $v_{c,max} \approx 140$ km s⁻¹. For this halo the Hernquist profile with parameters $m_s = 2.5 \times 10^{11} M_\odot$ and $r_s = 13$ kpc fits very well on scales up to ~ 20 kpc (although it appears to fall $\sim 20\%$ too low on larger scales). These parameters agree well with those obtained by Dubinski and Carlberg⁸, scaled to the virial mass of the halo, that is $m_s \propto v_{c,max}^3$, and assuming that $r_s \propto v_{c,max}$. The rotation curves presented by Warren *et al.*⁹ appear to demonstrate that this scaling holds down to the smallest haloes in their simulations, which have amplitude ~ 60 km s⁻¹.

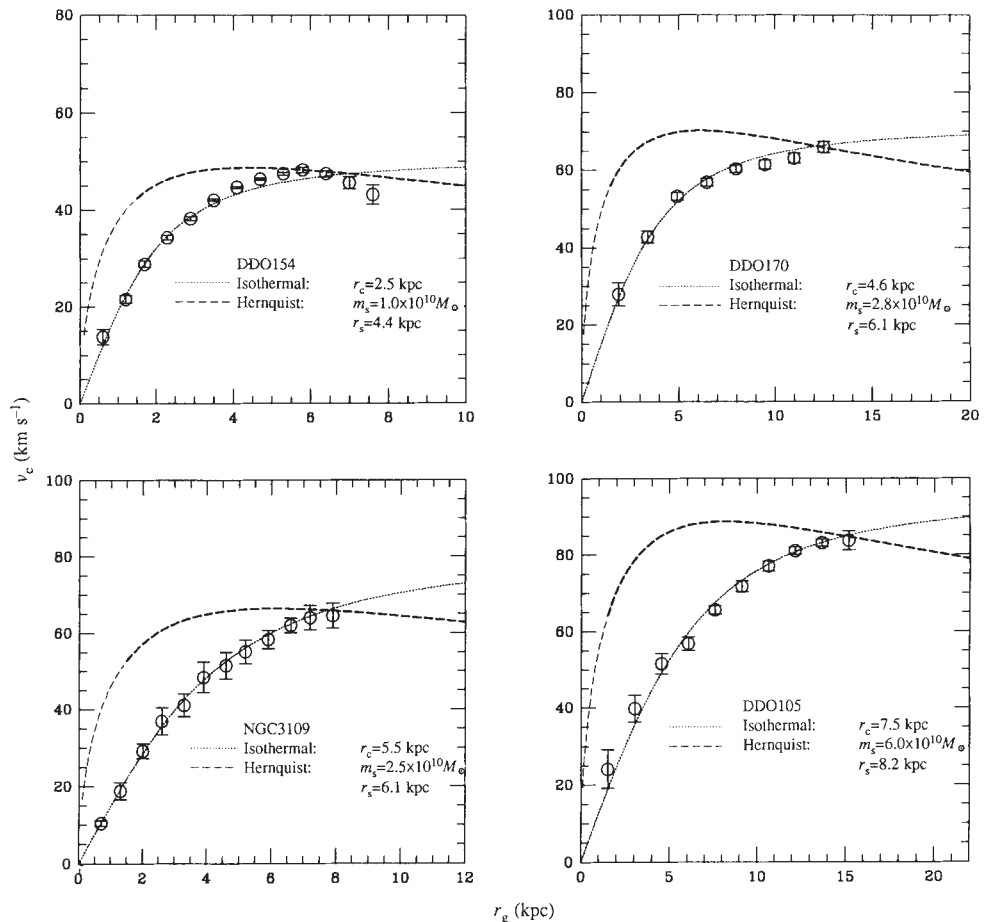
Determining the structure of dark haloes from the observations is difficult, and can be constrained only in a handful of galaxies because of uncertainties in the mass-to-light (M/L) ratio of the disk material. (The luminous mass—stars and molecular and atomic gas—can contribute a large fraction to the total mass within the central halo and must be subtracted to find the dark-matter distribution.) Typically, M/L ratios of the order of unity are adopted, which would imply that in a large spiral galaxy like our own the gravity of the luminous material dominates that of the dark matter to distances (from the centre) of the order of ~ 10 kpc (refs 20–22). On scales $\lesssim 10$ kpc, in this example the dark matter density is not falling as steeply as r_g^{-2} , and this core radius demarcates a transition between the regions where dark matter and luminous matter dominate. By arbitrarily increasing the disk M/L ratio the role of the dark halo can be reduced, at the expense of increasing the core radius. Lowering the disk M/L ratio reduces the core radius. Dynamical models of the disk yield values²³ of $M/L \approx 1$, however.

Dwarf spiral galaxies provide excellent probes of the internal structure of dark haloes, as these galaxies are completely dominated by dark matter on scales larger than a kiloparsec (refs 10–13). Furthermore, most of the baryonic mass is in the form of neutral hydrogen (H I) within the disk, the mass of which can be determined using the observed 21-cm emission. We can therefore use these galaxies to study the inner structure of dark haloes with very little ambiguity about the contribution from the luminous material and resultant uncertainties in the disk M/L ratio. Less than a dozen rotation curves have been measured for dwarf galaxies, but a trend is clearly apparent: the rotational velocity of the gas rises over most of the observed region, which goes many optical scale lengths into the dark halo, but which is within the core of the mass distribution (Fig. 1). For example, the rotation curve of the dwarf galaxy DDO154 is dominated by dark matter beyond ~ 1.5 kpc (ref. 10). In this system the mass of H I ($2.7 \times 10^8 M_\odot$), is one-fifth that of the stellar component, and the atomic gas extends over 15 optical scale lengths into the dark halo, which has total mass $\geq 5 \times 10^9 M_\odot$. Although dwarf spirals such as the DDO galaxies have luminosities only a few per cent of an L_* galaxy, they are the most numerous, and a large fraction of the mass in the Universe is associated with these systems.

Figure 1 shows the predicted rotation curves for several dwarf spiral galaxies, assuming that the dark haloes are dominated by

* Present address: Department of Astronomy, FM-20, University of Washington, Seattle, Washington 98195, USA.

FIG. 1 The observed rotation curves of four isolated dwarf spiral galaxies (data points); v_c is the rotation velocity of material at a distance r_g from the centre of each galaxy. The dotted lines show a fit to the observed rotation curve assuming zero contribution from the baryons, and a dark halo with an isothermal density profile $\rho(r_g) \propto 1/(r_c^2 + r_g^2)$, where r_c is the core radius. These 'maximum halo' models are designed to minimize the contribution of the disk, hence provide a lower limit to r_c , and they demonstrate that even if the gravity of the baryons is ignored, the dark haloes must have large core radii. If the observed baryons are taken into account, we find that r_c is actually $\sim 25\%$ larger. The core radii of these dark haloes are fairly tightly constrained and show an interesting trend of increasing size with $v_{c,\text{max}}$, scaling approximately as measured by Kormendy³³. (References for the data: DDO154 (ref. 11), DDO170 (ref. 12), NGC3109 (ref. 13), DDO105 (ref. 3).) The dashed lines show the rotation curves obtained adopting a Hernquist density profile with effective radius, r_s , scaled using $r_s \propto v_{c,\text{max}}$ (observed) and the effective mass, m_s , scaled so that the curve passes through this same point. These curves are drawn fainter on scales below the resolution limit of the simulations. The model predictions overestimate the observed mass within $r_c/2$ by a factor of 4 in each galaxy. The rotation curve of NGC3109 implies that for $r < 4$ kpc, $\rho(r) \propto r^{0.0}$ to 5% accuracy, whereas the Hernquist profile predicts $\rho(r) \propto r^{-1}$. If r_s and m_s are left as free parameters, values of r_s as large as 75 kpc are required



before an acceptable fit can be made to the rotation curves of dwarf galaxies³⁴.

collisionless particles. The model profiles are too steep on small scales, and clearly overestimate the observed rotational velocity for $r_g \lesssim 10$ kpc. The disparity between the models and the data follows directly from the nature of collisionless particles, which have no physically associated length scale, with the result that dark matter haloes should have zero effective core radii. This applies to any model in which the dark matter is cold and collisionless (for example, mixed dark-matter models²⁴), CDM models with a cosmological constant²⁵ or models with various combinations of cosmological density parameter and Hubble constant. Clearly, in order to give rise to cores in the dark-matter distribution, additional physical constraints are required, such as could be provided by HDM or an open universe dominated by baryonic dark matter. Although HDM may have been excluded from the dark matter candidate list for other reasons^{16,17}, phase space constraints do provide a natural mechanism that can limit the central density of dark-matter particles. Baryonic dark matter can interact with itself via electromagnetic forces as well as gravity; hence a characteristic scale can be introduced, perhaps resulting from supernova-driven winds or cooling flows.

Are there any observational biases which could lower the true velocities on small-scales and give rise to a false core in the mass distribution? In order to avoid problems introduced by telescope resolution, we have plotted data points only at separations equal to the half-power beam width (~ 20 – 40 arcsec). Beam smearing is important over scales larger than the beam width, and particularly over the region where the rotation curve is rising, which is typically over most of the data for dwarf galaxies. Detailed

modelling of the effects of beam smearing using both simulated and real data²⁶, shows that the velocities are biased low by $\lesssim 10\%$ over the rising part of the rotation curve, and as the curve flattens the effect becomes negligible. We note that several dwarf galaxies, such as NGC3109 (ref. 27), have been mapped at much higher resolution using optical Fabry–Perot techniques, which alleviate any remaining worries about the presence of false dark-matter cores in galaxy haloes. All of the galaxies in Fig. 1 have inclinations $> 60^\circ$, the optimum range for minimizing errors. The internal velocity dispersion of the gas is ~ 7 – 10 km s^{-1} , which provides a small amount of pressure support (corrected for in the data by using the asymmetric drift equation). This correction is typically $< 5\%$ of the observed rotational velocity, and is not biased towards lowering the velocities within the central halo^{22,28}.

Unless the numerical simulations are giving misleading results on scales 10 times their resolution limit, I conclude that the dominant dark-matter component cannot be composed of collisionless, weakly interacting particles (CDM). Numerical effects, such as two-body relaxation, appear to be unimportant, as identical results were obtained with different N -body codes and after increasing the numbers of particles by an order of magnitude: haloes in the N -body simulations are typically resolved by many thousands of particles. Although the numerical simulations discussed above follow only a dissipationless component, gas cools efficiently within dark haloes²⁹ and can modify the halo mass distribution. For example, the adiabatic cooling of gas, which contracts to form the bulge and disk, increases the central density of dark matter and would tend to erase any pre-existing core

radii^{30,31}. For a baryonic fraction of 10%, the final core radius of a dark halo may have contracted by up to a factor of three after the baryons have cooled³⁰, thus making the case against collisionless dark matter even stronger by reducing the size of a pre-existing core³². □

Received 21 January; accepted 11 July 1994.

1. Faber, S. M. & Gallagher, J. S. *A. Rev. Astr. Astrophys.* **17**, 135–187 (1979).
2. Ashman, K. M. *Publ. astr. Soc. Pacif.* **104**, 1109–1138 (1990).
3. Schramm, D. N. *Nucl. Phys. B. proc. suppl.* **28A**, 243–253 (1992).
4. Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M. *Astrophys. J.* **292**, 371–394 (1985).
5. White, S. D. M., Frenk, C. S., Davis, M. & Efstathiou, G. *Astrophys. J.* **313**, 505–516 (1987).
6. Moore, B., Frenk, C. S. & White, S. D. M. *Mon. Not. R. astr. Soc.* **261**, 827–846 (1993).
7. Bertschinger, E. *Astrophys. J. Suppl. Ser.* **58**, 39–66 (1985).
8. Dubinski, J. & Carlberg, R. *Astrophys. J.* **378**, 496–503 (1991).
9. Warren, S. W., Quinn, P. J., Salmon, J. K. & Zurek, H. W. *Astrophys. J.* **399**, 405–425 (1992).
10. Carignan, C. & Beaulieu, S. *Astrophys. J.* **347**, 760–770 (1989).
11. Lake, G., Schommer, R. A. & van Gorkom, J. H. *Astr. J.* **99**, 547–560 (1990).
12. Jobin, M. & Carignan, C. *Astr. J.* **100**, 648–662 (1990).
13. Broels, A. H. thesis, Univ. Groningen (1990).
14. Sarkar, S. *Observational Tests of Cosmological Inflation* (eds Shanks, T., Banday, A., Ellis, R. S., Frenk, C. S. & Wolfendale, A. W.) 91–102 (NATO Adv. Study Instit. No. 348, Kluwer Academic, Dordrecht, 1991).
15. White, S. D. M., Davis, M. & Frenk, C. S. *Mon. Not. R. astr. Soc.* **209**, 27P–31P (1984).
16. Gerhard, O. E. & Spergel, D. N. *Astrophys. J.* **389**, L9–L11 (1992).
17. Silk, J. & Vilenkin, A. *Phys. Rev. Lett.* **53**, 1700–1703 (1984).
18. Frenk, C. S., White, S. D. M., Efstathiou, G. P. & Davis, M. *Nature* **317**, 595–597 (1985).
19. Quinn, P. J., Salmon, J. K. & Zurek, W. H. *Nature* **322**, 329–335 (1986).
20. Athanassoula, E., Bosma, A. & Papaioannou, S. *Astr. Astrophys.* **179**, 23–40 (1987).
21. Kent, S. M. *Astr. J.* **93**, 816–832 (1987).
22. Lake, G. & Feinberg, L. *Astr. J.* **98**, 166–179 (1989).
23. Kuijken, K. & Gilmore, G. *Astrophys. J.* **367**, L9–L13 (1991).
24. Davis, M., Summers, F. J. & Schlegel, D. *Nature* **359**, 393–396 (1992).
25. Efstathiou, G. P., Bond, J. R. & White, S. D. M. *Mon. Not. R. astr. Soc.* **258**, 1P–6P (1992).
26. Begeman, K. G. *Astr. Astrophys.* **223**, 47–60 (1989).
27. Carignan, C. *Astrophys. J.* **299**, 59–73 (1985).
28. Skillman, E., Bothun, G., Murray, C. & Warmels, A. *Astr. Astrophys.* **185**, 61 (1987).
29. Katz, N., Hernquist, L. & Weinberg, D. H. *Astrophys. J.* **399**, L109–L112 (1992).
30. Blumenthal, G., Faber, S., Flores, G. & Primack, J. *Astrophys. J.* **301**, 27–34 (1986).
31. Carlberg, R. G., Lake, G. & Norman, C. A. *Astrophys. J.* **300**, L1–L4 (1986).
32. Flores, G. & Primack, J. *Astrophys. J.* (in the press).
33. Kormendy, J. in *Evolution of the Universe of Galaxies* (ed. Kron, R. G.) 33 (Astr. Soc. Pacific Vol. 10, Provo, Utah, 1990).
34. Sanders, R. H. & Bergman, K. G. *Mon. Not. R. astr. Soc.* **266**, 360 (1994).

ACKNOWLEDGEMENTS. I thank M. Davis for many useful and interesting discussions.

Multiphoton-induced X-ray emission at 4–5 keV from Xe atoms with multiple core vacancies

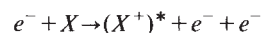
A. McPherson, B. D. Thompson, A. B. Borisov, K. Boyer & C. K. Rhodes

Department of Physics (M/C 273), University of Illinois at Chicago, 845 West Taylor Street, Chicago, Illinois 60607–7059, USA

SEVERAL recent experimental findings^{1–3} have pointed to a possible route for making an X-ray laser, which could in principle provide an imaging system capable of molecular resolution⁴. The method involves the multiphoton excitation of atoms in van der Waals clusters or in molecules to yield ions with core-electron vacancies^{1,2}, which can then decay by emission of X-rays, in conjunction with a self-channelling propagation mode of electromagnetic radiation³. The multiphoton excitation may be stimulated by ultrahigh-brightness, subpicosecond pulses of laser light⁵. We have previously observed² emission of X-rays from L-shell transitions in core-excited krypton atoms using this approach. Here we report the multiphoton production of X-rays of wavelength 2–3 Å from highly ionized xenon atoms which possess a large number of inner-shell vacancies while retaining several electrons in relatively weakly bound outer orbitals. Atoms with this ‘inverted’ electronic configuration are designated ‘hollow atoms’^{6,7}. We find that generation of hollow atoms can become the dominant excitation mode

for such systems, making their exploitation in an X-ray laser a real possibility.

In the cluster/molecule approach proposed previously for X-ray amplification, a molecule or a loosely bound cluster of atoms is irradiated by an intense ($>10^{16}$ W cm⁻²) source with a wavelength matched to the excitation conditions of the atomic constituents. An initially ionized free electron in the cluster or molecule may then lead to further ionization via intra-cluster inelastic electron-atom collisions:



By adjusting the excitation intensity depending on the inter-atomic spacings and inner-shell binding energies, processes such as these can lead to the selective generation of inner-shell excited species. We have tested some of the predictions of this model previously for the production of Kr(L) (5–7-Å) emission². A significant feature of the Kr(L) spectrum was the anomalous strength of the satellite lines, a property that indicates the occurrence of inner-shell ($n=2$) excitation while outer-shell ($n=3$) electrons remain bound to the atom.

We have now obtained experimental evidence for the direct multiphoton production in gas-phase xenon clusters of hollow atoms having multiple L-shell vacancies. Specifically, we have observed Xe(L) emission in the 2–3 Å range produced by excitation of the clusters with ultra-intense subpicosecond ultraviolet radiation. One measurement involves the recording of X-ray ‘pinhole’ CCD (charge-coupled device) camera images of the radiating zone with a spatial resolution of ~ 100 μm. Another entails spectral measurements obtained with a von Hámos curved crystal spectrometer⁸. Both studies were performed with the apparatus and methods previously described². The clusters

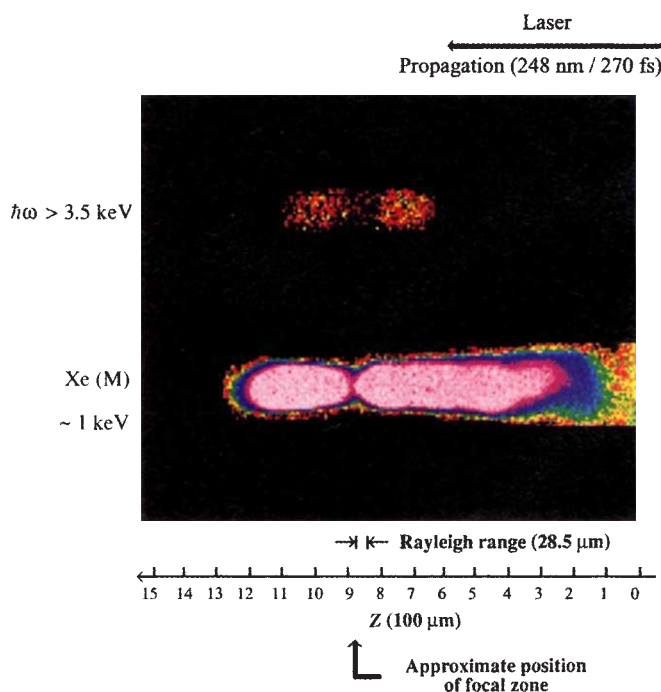


FIG. 1 X-ray images produced by a single exposure with the double pinhole camera. The upper and lower pinholes correspond to the Xe(L) and Xe(M) emissions, respectively. The pinholes had a diameter ~ 100 μm, limiting the spatial resolution to that value. The stagnation pressure of the pulsed jet was 75 pounds per square inch above atmospheric pressure and the temperature of the nozzle was 233 K. The approximate location of the focal region is indicated along with the corresponding Rayleigh range of the focusing system ($f/3$). The more restricted spatial distribution of the shorter wavelength emission ($\hbar\omega_x > 3.5$ keV) is expected, as significantly higher intensities are needed for the L-shell excitation relative to those required for the corresponding M-shell excitation^{1,2}. Intensity grey scale is presented in pseudocolour.

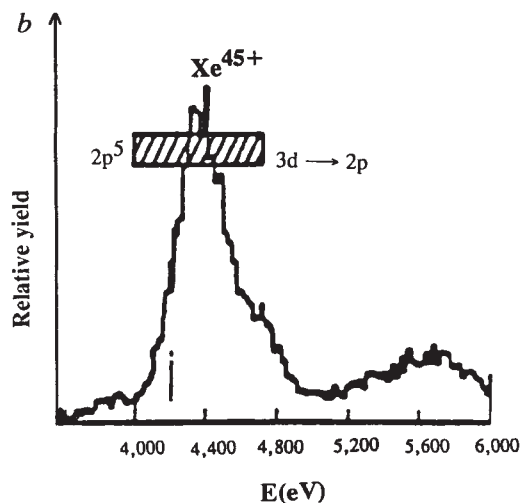
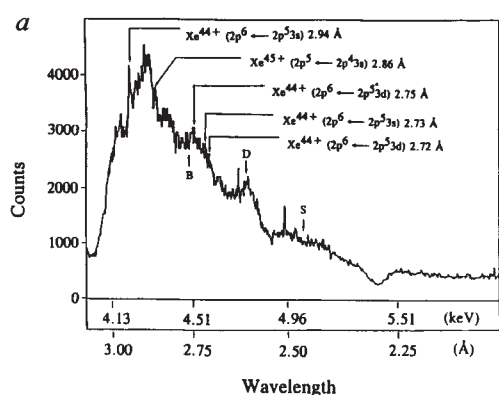
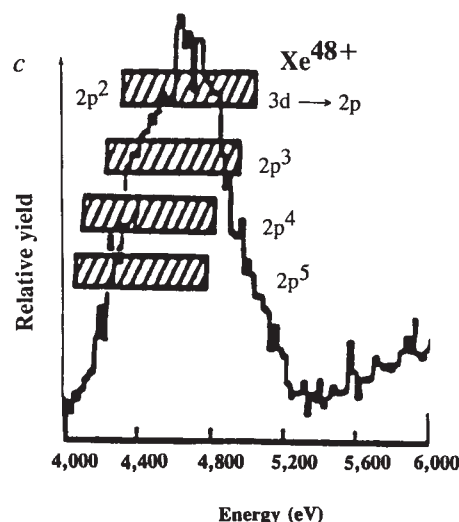


FIG. 2 a, The spectrum of the short-wavelength component ($\hbar\omega_x > 3.5$ keV) shown in Fig. 1, recorded with 720 laser pulses by a von Hámos spectrometer operating in third order. The spectrum shows a single broad feature having a width of ~ 700 eV and an asymmetrical wing extending to the blue side. Additional features are also indicated and identified. Although occurring generally in the expected spectral region for L-shell transitions in Xe, the measured spectrum is entirely different from that normally anticipated (for example, Ne-like). Specifically, the customary L-shell spectrum at a low average plasma density ($n_e \sim 10^{20} \text{ cm}^{-3}$) is dominated by a few sharp lines. In contrast, it is evident that the multiphoton-produced spectrum illustrated is predominantly composed of a very broad single feature. The film was read with a film scanning system (Materials Data Inc., Livermore, California) using an Agfa (Studio Scan) scanner for negatives. D denotes a localized defect in the film. The estimated accuracy of the abscissa is ± 50 eV. b, Xe(L) emission recorded in the 4–6 keV range following the impact of slow Xe^{45+} ions on Cu surfaces as described in ref. 7. The features in the 4–5 and 5–6 keV regions arise from $3d \rightarrow 2p$ and $4d \rightarrow 2p$ transitions, respectively, in an atom that possesses a large number of $n \geq 3$ electrons. On account of the interactions associated with the presence of the outer ($n \geq 3$) electrons, the allowed transitions become distributed over a large spectral region and a broad band of emission develops. The correspondence of the $3d \rightarrow 2p$ emission in the 4–5 keV range with the multiphoton-produced spectrum shown in a is marked. The abscissa has been adjusted to closely match that of a. The estimated accuracy of the abscissa is ± 50 eV. (Shown with permission⁷.) c, Corresponding Xe(L) emission recorded following the impact of slow Xe^{48+} ions on a Cu surface⁷ which also manifests properties similar to the spectrum in a. The $2p^2$ and $2p^3$ (shaded) zones overlap with a shoulder at ~ 5 keV



and feature S in a. The estimated accuracy at the abscissa is ± 50 eV. (Shown with permission⁷).

were produced in a cooled pulsed-gas jet. We use a laser⁵ with a wavelength of 248 nm, a pulse length of ~ 300 femtoseconds (fs, 10^{-15} s) and a peak power of ~ 0.7 terawatts (TW, 10^{12} W). The focusing system produces a maximum intensity of ~ 8 exawatts cm^{-2} (EW, 10^{18} W).

The X-ray 'pinhole' camera pictures are single-shot exposures obtained by viewing, in a direction transverse to the propagation of the 248-nm beam, the plasma excited by the focused subpicosecond 248-nm radiation directly below the pulsed jet. The camera has two pinholes that can record images simultaneously. One pinhole has a stainless-steel filter with a thickness of 12.7 μm , making it entirely 'blind' to radiation with a quantum energy $\hbar\omega_x < 3.5$ keV. The other pinhole has a filter with the composition (2,000 Å Al + 1 μm polycarbonate + 10 μm Be) so that all radiation with a quantum energy $\hbar\omega_x \geq 700$ eV is transmitted. Furthermore, as the Xe(M) emission at ~ 1 keV was found experimentally to be considerably brighter than the Xe(L) radiation at ~ 4 keV, with the use of these filters, one image represents the spatial distribution of the Xe(M) radiation while the other presents the corresponding distribution of the Xe(L) emission. The experimentally observed spatial distributions are shown in Fig. 1.

In Fig. 2 we show the spectra in the 4–5 keV range obtained by our method (Fig. 2a), and those for Xe(L) emission following the impact of Xe^{45+} (Fig. 2b) and Xe^{48+} (Fig. 2c) ions on a copper surface⁷, which were shown to come from hollow-atom

states. The spectra share several characteristics including general spectral position, breadth and particular features. The overall correspondence leads to the conclusion that the $3d \rightarrow 2p$ emissions observed in both cases arise from the same type of ionic states, namely hollow-atom configurations in which inner-shell vacancies are associated with the presence of a large number (~ 10) of outer-shell ($n \geq 3$) electrons.

The multiphoton spectrum shown in Fig. 2a has a clear asymmetry, weak narrow components superimposed on the dominant broad feature, and visible modulations in the intensity that can be correlated with spectral characteristics observed in studies of ion-surface emission^{6,7}, laser-imploded microballoons⁹ and radiatively driven capsules¹⁰. The asymmetry of the multiphoton spectrum can be explained by the presence of multiple L-shell vacancies. For example, the broad feature S in Fig. 2a at ~ 5 keV lies at a position representing a null in Fig. 2b, but this region corresponds to a significant shoulder in Fig. 2c which overlaps the multiple-vacancy $2p^2$ and $2p^3$ zones (shaded). Three transitions in Xe^{44+} (2.72–2.75 Å), whose known wavelengths are indicated to coincide with the feature B, correspond to the strongest lines observed in the microballoon⁹ and capsule¹⁰ studies. Additionally, two narrow transitions are identified in the multiphoton spectrum at 2.94 and 2.86 Å, attributed to $3s \rightarrow 2p$ transitions in Xe^{44+} and Xe^{45+} , respectively. The clear appearance of the Ne-like Xe^{44+} ($3s \rightarrow 2p$) line mirrors the observation of the corre-

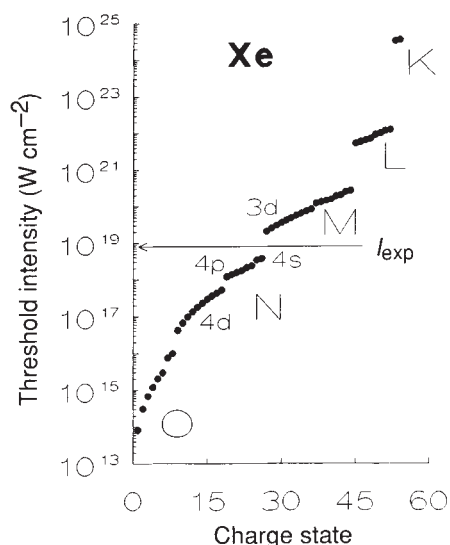


FIG. 3 Threshold intensities for ionization of electrons from bound states of xenon as derived from ref. 12. Specific electron shells and levels are indicated. With a maximum experimental intensity $I_{\text{exp}} \approx 8 \times 10^{18} \text{ W cm}^{-2}$, the Coulomb-suppression picture leads simultaneously to the disappearance of 4d states and the retention of the 3d electrons. Hence, the absence in the multiphoton spectrum illustrated in Fig. 2a of the 4d→2p band shown in Fig. 2b and c is another feature of experimental agreement between the results of the multiphoton excitation, the findings of the ion-surface studies⁷ and the expectations of the cluster approach^{1,2}.

sponding Ne-like Kr^{26+} ($3s \rightarrow 2p$) transition as the strongest feature in multiphoton produced Kr(L) emission².

The presence of the Xe^{45+} species with the excited-state configuration ($2p^4 3s$) and the overlap of the feature S with the spectrum shown in Fig. 2c arising from the Xe^{48+} ions confirm the direct production of double vacancies by the multiphoton coupling and suggest that higher levels of 2p shell ionization are also present. The generation of the Xe^{45+} ($2p^4 3s$) state from the neutral atom requires an energy transfer¹¹ exceeding 70 keV.

The multiphoton spectrum in Fig. 2a does, however, exhibit a significant difference from those produced by ion-surface collisions⁷: the feature at ~5.5 keV in Fig. 2b and c, which corresponds to the 4d→2p component, is absent in Fig. 2a. This fact

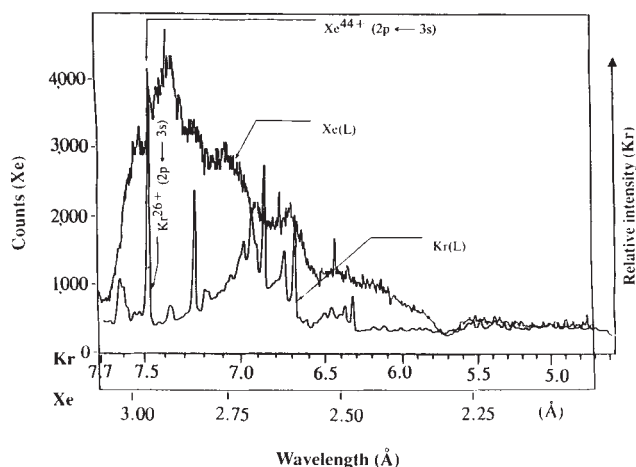


FIG. 4 Comparison of the morphologies of the multiphoton-produced emission for Kr(L) (ref. 2) and Xe(L). The two spectra are registered by the corresponding ($2p \leftarrow 3s$) Ne-like transitions for Kr^{26+} and Xe^{44+} . The conversion from a line-dominated spectrum for Kr to a continuous distribution reflecting the characteristics of the hollow-atom configurations produced with Xe is clear.

is easily explained by the Coulomb-suppression model of multiphoton ionization¹², the threshold conditions for which are illustrated in Fig. 3.

The comparison of the morphologies of the multiphoton-produced Kr(L) and Xe(L) spectra shown in Fig. 4 is informative. If the spectra are registered to the corresponding $2p \leftarrow 3s$ Ne-like Kr^{26+} and Xe^{44+} features, the two spectra are essentially isomorphic with the major exception that the gaps between the dominant sharp features of the Kr(L) distribution are filled in by the high density of satellite lines seen in the Xe(L) emission. Hence the line spectrum for charge state $Z=36$ is replaced at $Z=54$ by satellite transitions characteristic of the hollow-atom spectrum. Extrapolation of this trend leads to the conclusion that, for heavy atoms in clusters and appropriately designed molecules, the multiphoton excitation of hollow-atom states will be the predominant excitation.

A comparison of the spectrum shown in Fig. 2a with a spectrum thermally produced¹⁰ with an electron temperature $T_e \lesssim 1.4 \text{ keV}$ at a comparable electron density ($n_e \approx 10^{24} \text{ cm}^{-3}$), as shown in Fig. 5, reveals a significant difference. The thermal spectrum deviates grossly from the multiphoton-generated distribution. Specifically, the former is essentially null between 2.8 and 3.0 Å, the region corresponding to the strong maximum in the latter. Conversely, the high degree of spectral correspondence shown between Fig. 2a, b and c, as the effective electron temperature associated with the ion-surface interaction is very low ($\lesssim 5 \text{ eV}$), indicates that the multiphoton-produced emission involves a negligible contribution from thermal excitation. Because the multiphoton spectrum must involve electron colli-

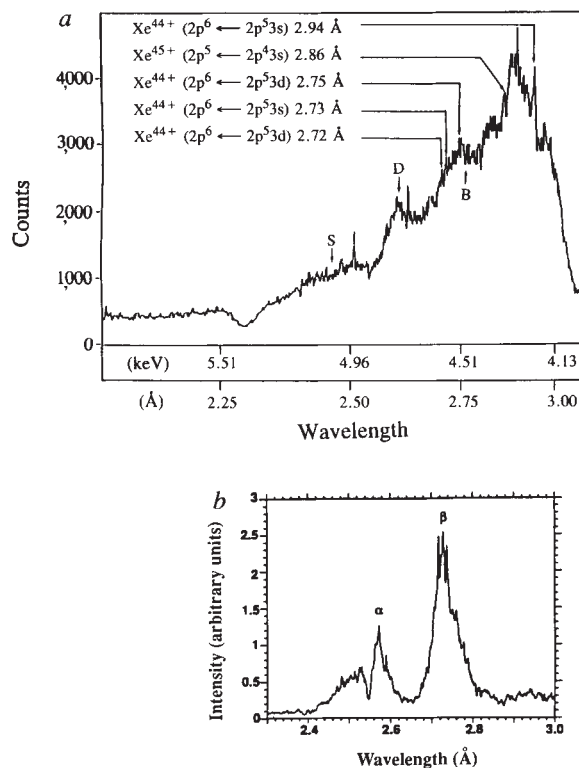


FIG. 5 Comparison of the multiphoton-generated Xe(L) spectrum (a) with a thermally produced spectrum¹⁰ (b) obtained at a comparable electron density ($n_e \approx 10^{24} \text{ cm}^{-3}$). The thermal spectrum was obtained at a temperature estimated¹⁰ to be $T_e \lesssim 1.4 \text{ keV}$. In contrast to the spectral similarity exhibited in Fig. 2, these two spectra correspond poorly, particularly in the important region where the maximum of the multiphoton spectrum occurs (~2.8–3.0 Å). (b shown with permission¹⁰). S is a significant shoulder feature, D is a localized film defect, B is a strong line (see text), α is an identified strong feature of the thermal spectrum¹⁰ and β is a feature of the thermal spectrum corresponding to B.

sions in the kilovolt range¹, this signifies that electron motions produced by the external field occupy only a very small fraction of the available phase space. Therefore, we can conclude that the electronic motions causing the observed multiphoton excitation possess a very significant level of order and are not appreciably perturbed by a thermal component^{13–15}. □

Received 14 February; accepted 11 July 1994.

- McPherson, A., Luk, T. S., Thompson, B. D., Boyer, K. & Rhodes, C. K. *Appl. Phys.* **B57**, 337–347 (1993).
- McPherson, A. et al. *Phys. Rev. Lett.* **72**, 1810–1813 (1994).
- Borisov, A. B. et al. *Phys. Rev. Lett.* **68**, 2309–2312 (1992).
- Solem, J. C. & Baldwin, G. C. *Science* **218**, 229–235 (1982).
- Bouma, B., Luk, T. S., & Rhodes, C. K. *J. opt. Soc. Am.* **B10**, 1180–1184 (1993).
- Schneider, D. et al. *Phys. Rev.* **A42**, 3889–3895 (1990).
- Clark, M. W. et al. *Phys. Rev.* **A47**, 3983–3997 (1993).
- Mazing, M. A. & Shevel'ko, A. P. *Proceedings of the Lebedev Physical Institute* Vol. 179 (ed. Sobel'man, I. I.) 1–50 (Nova Science, Commack, New York, 1988).
- Contunlie, Y., Yaakobi, B., Feldman, U., Daschek, G. A. & Cowan, R. D. *J. opt. Soc. Am.* **71**, 1309–1314 (1981).
- Keane, C. J., Hammel, B. A., Osterheld, A. L. & Kania, D. R. *Phys. Rev. Lett.* **72**, 3029–3032 (1994).
- Carlson, T. A., Nestor, C. W. Jr, Wasserman, N. & McDowell, J. D. *Atomic Data* **2**, 63–99 (1970).
- Augst, S., Strickland, D., Meyerhofer, D. D., Chin, S. L. & Eberly, J. A. *Phys. Rev. Lett.* **63**, 2212–2215 (1989).
- Boyer, K., Thompson, B. D., McPherson, A. & Rhodes, C. K. *J. Phys. B* (in the press).
- Thompson, B. D., McPherson, A., Boyer, K. & Rhodes, C. K. *J. Phys. B* (in the press).
- Boyer, K. & Rhodes, C. K. *J. Phys. B* (in the press).

ACKNOWLEDGEMENTS. The authors thank J. Wright and P. Noel for technical assistance, and S. Guggenheim for the use of his microdensitometer, and A. P. Shevel'ko for measurements with the von Hámos spectrometer. C.K.R. acknowledges discussions with D. Schneider and J. W. Longworth. This work was supported by the US DoE, AFOSR, ONR, SDI/NRL and ARO.

Reduction of frictional forces between solid surfaces bearing polymer brushes

Jacob Klein*, Eugenia Kumacheva*,
Diana Mahalu*, Dvora Perahia*
& Lewis J. Fetters†

* Department of Materials and Interfaces,
Weizmann Institute of Science, Rehovot 76100, Israel
† Exxon Research and Engineering Company, Annadale,
New Jersey 08801, USA

THE use of lubricants to reduce friction and wear between rubbing surfaces has been documented since antiquity^{1–3}. Recent approaches have focused on boundary lubrication by surfactant-like species coating the surfaces, whereby the friction between them is replaced by the weaker forces required for shear of adhesive contacts between the surfactant layers^{3,4}. An alternative approach is to tether polymer chains to the surfaces by one end which, when swollen by a solvent, then act as molecular 'brushes' that may facilitate sliding. The normal forces between sliding brush-bearing surfaces have been previously investigated^{5,6}, but the lateral forces, which are the most important from the point of view of lubrication, are harder to measure. Here we report the measurement of lateral forces in such a system. We find a striking reduction in the effective friction coefficients μ_b between the surfaces to below our detection limit ($\mu_b < 0.001$), for contact pressures of around 1 MPa and sliding velocities from zero to 450 nm s⁻¹. We believe that this effect is due to the long-ranged repulsion, of entropic origin, between the brushes, which acts to keep the surfaces apart while maintaining a relatively fluid layer at the interface between them.

Normal and lateral forces, $F_{\text{normal}}(D)$ and $F_{\text{shear}}(D)$ respectively, between curved, molecularly smooth mica surfaces immersed in toluene were determined as a function of the surface separation D and the sliding velocity v_s . The apparatus used, described in detail earlier^{5,6}, is capable of measuring the forces

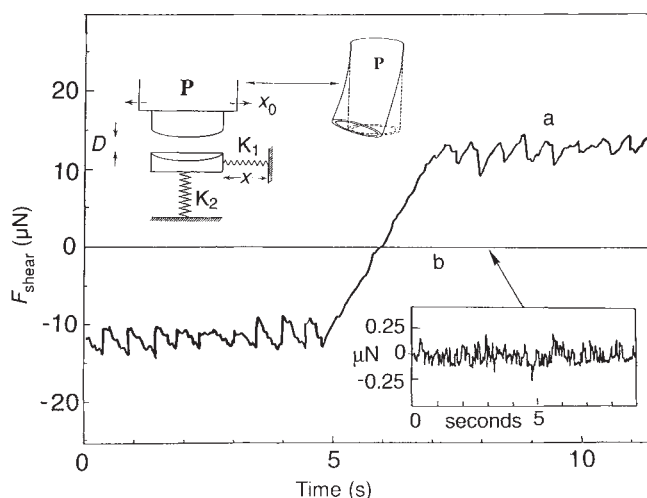


FIG. 1 The variation with time of the shear force F_{shear} between two mica sheets a distance D apart, compressed by a normal load $F_{\text{normal}} = 19 \pm 2 \mu\text{N}$, in response to an applied shear. The upper inset shows schematically the configuration of the surfaces mounted on crossed cylindrical lenses. The bending δx and δD of the two orthogonal springs K_1 and K_2 yields the respective forces F_{shear} and F_{normal} (ref. 5). Lateral motion δx_0 is applied to the top surface via the sectorized piezocrystal P, whose bending configuration is indicated. The sliding velocities used in the present study were in the range $v_s \equiv d(x_0 - x)/dt = 15\text{--}450 \text{ nm s}^{-1}$. Curve **a**, sliding of bare mica surfaces in toluene. $D = 14 \pm 3 \text{ \AA}$, corresponding to two monolayers of toluene between the surfaces. A clear stick-slip behaviour is observed during sliding. The region where the surfaces stick and the shear force changes sign at $t = 5\text{--}7 \text{ s}$ indicates a reversal of the sliding direction. The sliding velocity $v_s = 45 \text{ nm s}^{-1}$. Curve **b**, Sliding of mica surfaces covered by a PS-X(1.4×10^5) brush, at $D = 37 \text{ nm}$. (In this and subsequent data the brushes form from a solution of polymer concentration $(6 \times 10^{-5})\text{--}(10^{-4}) \text{ g ml}^{-1}$. On the scale of the main figure the shear forces are within the thickness of the line labelled **b**. The lower inset shows F_{shear} on an expanded scale: the shear force is less than the noise-limited resolution δF_{shear} . There is no measurable change on reversing the sliding direction. The sliding velocity $v_s = 15 \text{ nm s}^{-1}$.

directly and simultaneously by monitoring the bending of two sets of springs as the surfaces move with respect to each other (Fig. 1, upper inset).

When the mica sheets are compressed in the pure solvent the toluene is squeezed out, until the normal force is balanced by the structural forces due to the final monolayers of solvent remaining between the surfaces, as observed for other liquids⁷. On applying a shear force that exceeds the frictional force, the surfaces slide and their motion displays a characteristic stick-slip behaviour⁸, as seen clearly in Fig. 1, curve **a**. Such stick-slip motion has been observed also for mica sheets in other simple liquids⁸. It is in accord with the model of interfacial sliding due to Tabor⁴, where the friction F arises from the forces required to shear the adhesive junctions: $F = S_c A$; here S_c is a

TABLE 1 Molecular characteristics of polymers and corresponding brushes

Sample	M_r	M_r/M_n	$2L_M(\text{nm})$	$s_M(\text{nm})$
PS-X(2.6×10^4)	2.65×10^4	1.02	45	4.6
PS-X(1.4×10^5)	1.41×10^5	1.03	125	9.4
PS-X(3.75×10^5)	3.75×10^5	1.03	230	14.4

Samples were synthesised and characterized as described in ref. 9. The end-group -X is $-(\text{CH}_2)_3\text{N}^+(\text{CH}_3)_2(\text{CH}_2)_3\text{SO}_3^-$. M_r is the relative weight-averaged molecular mass of the polymer, M_n is its relative number-averaged molecular mass and L_M and s_M are the brush heights and mean inter-anchor spacings determined from the respective $F_{\text{normal}}(D)$ profiles (the surface number densities of chains for each brush are given by $(1/s_M)^2$).

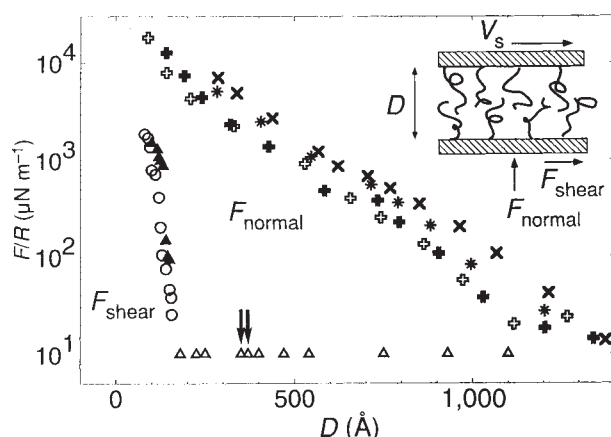


FIG. 2 Normal and shear force profiles, $F_{\text{normal}}(D)$ and $F_{\text{shear}}(D)$ respectively, between mica surfaces bearing PS-X(1.4×10^5) brushes, as indicated in the diagram. The force axis is normalized with respect to the radii of curvature R (~ 1 cm) of the mica sheets, which converts $F_{\text{normal}}(D)$ to the interaction energy in the Derjaguin approximation (for comparison F_{shear} is similarly normalized). Results from several different experiments are shown. Shear velocities for the $F_{\text{shear}}(D)$ measurements were in the range $v_s = 15$ – 450 nm s $^{-1}$. The open triangles (Δ) at $D > 180$ Å correspond to shear forces that are within the noise-limited resolution δF_{shear} of the signal (see Fig. 1, lower inset): as an upper estimate they are all set to equal $|\delta F_{\text{shear}}|$. Highlighted F_{shear} data at $D = 374$ Å and 352 Å (indicated by arrows) were taken at $v_s = 15$ nm s $^{-1}$ and 450 nm s $^{-1}$ respectively.

critical shear stress which depends on the details of the interfacial region, and A is the contact area.

Following measurements in pure toluene, end-functionalized polystyrene (PS) chains of relative molecular mass M_r , designated PS-X(M_r), where the end-group -X denotes a zwitterion, are added to the solvent. Their characteristics are given in Table 1. The molecules attach spontaneously by their zwitterion-terminated end to the mica surfaces, with the PS chain dangling into the solution⁹, to form extended brush-like layers whose structure and normal interactions have been extensively studied^{9,12}. Figure 1 curve *b* shows the shear forces required to slide surfaces covered by a PS-X(1.4×10^5) brush at the same normal load as for the pure toluene case. These are some orders of magnitude lower than the forces needed for sliding the bare surfaces, and indeed are within the vibration-limited resolution $\delta F_{\text{shear}} \approx \pm 0.1$ μN of the measurement (Fig. 1, lower inset).

The $F_{\text{normal}}(D)$ and $F_{\text{shear}}(D)$ profiles between the brush-bearing surfaces are shown in Fig. 2 for PS-X(1.4×10^5) over the entire range $D < 2L$ of interactions (where $L \approx 65$ nm is the thickness of each brush). $F_{\text{normal}}(D)$ is repulsive due to osmotic interactions; it increases monotonically at $D < 2L$, and is similar to earlier studies⁹. The corresponding shear forces required to induce and maintain interfacial sliding are extremely low. $F_{\text{shear}}(D)$ is within the instrumental resolution δF_{shear} down to $D = D_0 \approx 20$ nm; that is, under normal loads up to $F_{\text{normal}}(D = D_0) \approx 100$ μN. Defining an effective friction coefficient between the sliding brush-covered surfaces as $\mu_b = (F_{\text{shear}}(D)/F_{\text{normal}}(D))$, we find $\mu_b < 0.001$ (that is, it is smaller than our measurable limit) at these loads. In this regime the behaviour is reversible with respect to surface separation, indicating stability of the brushes to shear. For the regime $2L < D < D_0$, the frictional forces remain immeasurably small for all velocities in the range $v_s = 15$ – 450 nm s $^{-1}$; no frictional force was detectable when the surfaces started moving from rest. Clearly, whatever frictional force remains will be the result of viscous dissipation, and thus velocity dependent—but as it is below our instrumental resolution we cannot quantify this dependence. At higher compressions a stick-slip behaviour similar to that seen in the

absence of polymer is observed (Fig. 3 inset), with a measurable $F_{\text{shear}}(D)$ which increases rapidly at $D < D_0$ (at the highest compressions there was some indication that the surface-attached polymer was sheared off as the surfaces slid past each other, but following a short period of separation in the PS-X solution, the brushes 'healed' to their original form).

Similar profiles were measured between mica surfaces covered by PS-X(M_r) brushes with $M_r = 2.6 \times 10^4$ and 3.8×10^5 . This range of M_r corresponds to a fivefold variation in brush-thickness and tenfold variation in tethered chain density (Table 1). The data is shown in Fig. 3. In this reduced representation, a similar general behaviour is seen for all three brushes: extremely low friction (within the resolution δF_{shear} of the measurements) from $D = 2L$ down to a surface separation D_0 , and a rapidly increasing frictional force at higher compressions ($D < D_0$). We find that the onset for rapid increase in the friction is at a compression ratio $(D_0/2L) \approx 0.1$ – 0.15 for all three brushes. The ratios of shear to normal forces at $D = D_0$ yield the effective friction coefficients μ_b at these separations, and show that $\mu_b < 0.001$ in this regime for all three cases. The corresponding contact pressures P at the points of closest approach is given by ref. 9 as $P = (1/2\pi R)(\partial F_{\text{normal}}/\partial D)|_{D=D_0}$. This yields values $P(D_0) \approx (1-2) \times 10^6$ N m $^{-2}$, with the higher values for the shorter brushes.

The reduction in the frictional forces afforded by the layers of end-attached polymer chains is of entropic origin. Excluded-volume effects, which reduce the configurational entropy of the chains, dominate the interactions and lead to strong net segmental repulsions in the good-solvent medium^{13,14}. The result is a long-range equilibrium separation of the surfaces⁹⁻¹¹. At the same time theoretical studies suggest only a limited mutual interpenetration of the brushes^{15,16}, even under compression. This in turn ensures rapid relaxation of the interpenetrating chains as the surfaces slide past each other, and thus a relatively fluid region at the brush-brush interface: hence the low friction. The shear force F_{shear} required to maintain sliding of the brushes arises from the viscous dissipation in the interfacial region asso-

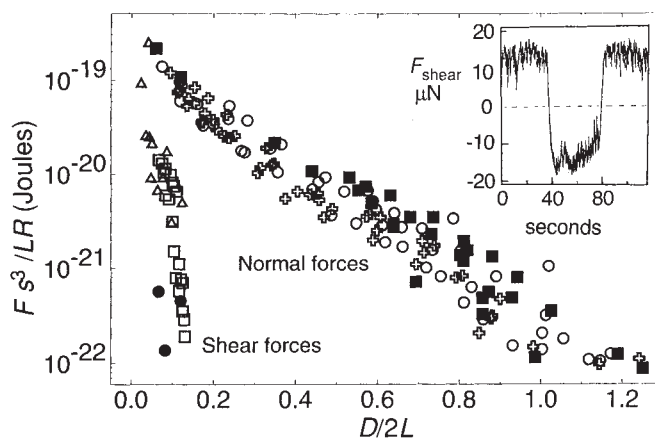


FIG. 3 Normal and shear force profiles for interactions between mica surfaces bearing PS-X(M_r) brushes, plotted as (F_s^3/LR) versus $D/2L$, where $s = s_M$ is the mean spacing between chain anchoring points for the respective brushes, and $L = L_M$ are the respective brush heights. In the Alexander-de Gennes model^{10,11} this reduction makes the F_{normal} data independent of M_r . The $F_{\text{shear}}(D)$ profiles have been similarly scaled for purposes of comparison. $\blacksquare$, $\bullet$, $M_r = 2.6 \times 10^4$; $\square$, $\circ$, $M_r = 1.4 \times 10^5$; Δ , $\oplus$, $M_r = 3.75 \times 10^5$. Results are shown from two or three experiments for each molecular weight; shear forces smaller than the resolution δF_{shear} are not shown. The inset shows a typical $F_{\text{shear}}(D)$ versus time plot for the highly-compressed ($D < D_0$) region of the plot, for PS-X(3.75×10^5) brushes compressed to surface separation $D = 190$ Å (corresponding to $D/2L = 0.083$); sliding velocity 38 nm s $^{-1}$; the change in sign of F_{shear} at 38 and 76 s corresponds to a reversal of the sliding direction at these times.

ciated with this mutual intertanglement; but as noted earlier, in the regime of low friction $2L < D < D_0$ any changes in F_{shear} with v_s were below the resolution δF_{shear} of our measurements.

At the highest compressions sliding friction increases as these relaxations become sluggish. Partly this is because the effect of mutual intertanglement of the brushes increases at higher concentrations; but mainly, we believe, it is due to divergence of the relaxation times as the polystyrene in the gap approaches a glassy phase. (Polymer films only a few nanometres thick can have a glass transition temperature similar to that in the bulk¹⁷.) The concentration c of polymer in a brush layer may be evaluated from the brush height L and its surface density (Table 1). For an uncompressed PS-X(1.4×10^5) brush (say), $c \approx 4.5\%$ w/v and varies inversely with D at $D < 2L$. For a compression ratio $(D/2L) = 0.1$, $c \approx 45\%$. The viscosity $\eta(c)$ of bulk polystyrene solutions at such high concentrations is high, and varies extremely rapidly with c . Typically, an increase in concentration from 45% to 55% in a solution of polystyrene, similar in size to the PS-X(1.4×10^5) chains, increases $\eta(c)$ by nearly an order of magnitude^{18,19}. For PS-X(1.4×10^5) brushes, such an increase in c requires only a small change in $(D/2L)$: 0.1 to 0.08. Similar considerations apply to the other brushes. Although the shear mechanism of chains in a brush and in bulk polystyrene solutions differs^{15,16}, it is likely that the relative increase in their viscosity—and thus in viscous dissipation—with c is comparable¹⁹. This is consistent with the very sharp rise in the shear force as D decreases at $D < D_0$. Brushes consisting of polymers with low glass transition temperatures should thus be better lubricants at high compressions.

The precise frictional mechanism is very different to the classic interfacial sliding picture of Tabor and Bowden^{4,20} and remains to be elucidated. Our observations reveal that a very fluid layer can reside at the interface between two solvated polymer brushes even under appreciable mutual compression, and can profoundly reduce the forces required for interfacial sliding of the surfaces bearing them. These results offer insight into relaxation processes in confined polymer layers and into tribological processes involving macromolecular surface phases. They also shed light on the origin of the very low friction in biological joints^{21,22}, where polymer-covered surfaces frequently rub against each other. Together with the self-assembling, robust and self-healing nature of polymer brushes, our results have clear implications for new lubrication strategies. □

Received 24 February; accepted 7 July 1994.

1. Tabor, D. *Friction* (Doubleday, New York, 1973).
2. Dowson, D. *History of Tribology* (Longmans, London, 1979).
3. Singer, I. L. & Pollock, H. M. (eds) *Fundamentals of Friction: Microscopic and Macroscopic Processes* (Kluwer, Dordrecht, 1992).
4. Tabor, D. in *Microscopic aspects of adhesion and lubrication* (ed. Georges, J. M.) 651–682 (Elsevier, New York, 1982).
5. Klein, J., Perahia, D. & Warburg, S. *Nature* **352**, 143–145 (1991).
6. Klein, J. *Pure appl. Chem.* **64**, 1577–1584 (1992).
7. Christenson, H. K. *J. chem. Phys.* **78**, 6906–6913 (1983).
8. Gee, M. L., McGuigan, P. M., Israelachvili, J. N. & Homola, A. M. *J. chem. Phys.* **93**, 1895–1906 (1990).
9. Taunton, H. J., Toprakcioglu, C., Fetters, L. J. & Klein, J. *Macromolecules* **23**, 571–580 (1990).
10. Alexander, S. *J. Phys. Paris* **38**, 983–989 (1977).
11. de Gennes, P. G. *Adv. Colloid Interface Sci.* **27**, 189–207 (1987).
12. Halperin, A., Tirrell, M. & Lodge, T. *Adv. Polymer Sci.* **100**, 31–96 (1991).
13. Flory, P. J. *Principles of Polymer Chemistry* (Cornell Univ. Press, Ithaca, 1953).
14. de Gennes, P. G. in *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, 1975).
15. Witten, T., Leibler, L. & Pincus, P. *Macromolecules* **23**, 824–830 (1990).
16. Joanny, J.-F. *Langmuir* **8**, 989–995 (1992).
17. Jackson, C. L. & McKenna, G. B. *J. non-cryst. Solids* **131–133**, 221–224 (1991).
18. Graessley, W. W., Hazleton, R. L. & Lindeman, L. R. *Trans. Soc. Rheology* **11**, 267–285 (1967).
19. Utracki, L. A. & Roovers, J. E. L. *Macromolecules* **6**, 366–372 (1973).
20. Bowden, F. P. & Tabor, D. *The Friction and Lubrication of Solids* (Clarendon, Oxford, 1950).
21. McCutchen, C. W. *Nature* **184**, 1284–1286 (1959).
22. Maroudas, A. in *Lubrication and Wear in Joints* (ed. Wright, V.) Ch. 14. (Sector, London, 1969).

ACKNOWLEDGEMENTS. We thank M. Tirrell, A. Maroudas, P. Pincus, S. Granick and particularly D. Tabor for discussions. Support from the Israeli Ministry of Science, the Juulich Forschungszentrum, the Minerva Foundation, the Levin Fund and the Israeli National Science Foundation are acknowledged with thanks. J.F. is the Herman Mark Chair of Polymer Physics at the Weizmann Institute.

Polymeric fullerene chains in RbC₆₀ and KC₆₀

Peter W. Stephens*, G. Bortel†, G. Faigel†, M. Tegze†, A. Jánosy†, S. Pekker†, G. Oszlanyi†§ & L. Forró§

* Department of Physics, State University of New York, Stony Brook, New York 11794, USA

† Research Institute for Solid State Physics, H-1525 Budapest, Hungary

‡ Technical University Budapest, Institute of Physics, H-1521 Budapest, Hungary

§ Laboratoire de Physique des Solides Semicristallins, IGA, Département de Physique, Ecole Polytechnique Fédérale de Lausanne, 1015-Lausanne, Switzerland

NEARLY all of the molecular crystals containing C₆₀ formed at ambient pressure^{1,2} have inter-fullerene separations of the order of 10 Å—the expected distance based on the molecular van der Waals radii. The sole exceptions are the room-temperature phases of AC₆₀ (where A denotes K, Rb or Cs), which are formed by reversible solid-state transformation from high-temperature (>150 °C) phases³. These phases have lattice parameters about 9% shorter in one direction, and in addition RbC₆₀ has magnetic properties suggestive of a one-dimensional metal⁴. We suggested in ref. 4 that this short distance may be due to covalent bonding between neighbouring C₆₀ molecules. Here we provide direct evidence for such bonding from powder X-ray diffraction studies of RbC₆₀ and KC₆₀. The linkage is through a [2 + 2] cycloaddition, which has been hypothesized to take place during photopolymerization of solid C₆₀ (ref. 5), and which has also been proposed⁶ for RbC₆₀. Such inter-fullerene linkages are calculated^{7,8} to be the preferred mode of dimerization of C₆₀. The AC₆₀ phases thus provide an example of a thermal phase transition driven by the reversible formation and breaking of covalent bonds.

TABLE 1 Results of fits to various structural models for RbC₆₀ and KC₆₀

No.	Details	RbC ₆₀			KC ₆₀		
		p	R _{WP}	F vs F ₉₅	p	R _{WP}	F vs F ₉₅
1.	Undistorted C ₆₀ , aligned with axes	13	13.7		22	12.3	
2.	Undistorted C ₆₀ , rotated ±45° about 9.11 Å direction	14	9.17	75 >> 4.0	23	9.95	25 >> 4.0
3.	Atoms C1 at inter-fullerene contact move in a direction.	15	6.19	72 >> 4.0	24	7.43	36 >> 4.0
4.	Atoms C1 move in molecular mirror plane	16	5.20	24.6 >> 4.0	25	6.41	15 >> 4.0
5.	Atoms C2 and C3 allowed to move, and overall C ₆₀ radius	20	4.85	8.2 > 2.5	29	5.78	2.4 < 2.6
6.	Same as 5, in Immm, with 50% C occupancy	20	4.93		29	6.19	

The fits shown here employ Rietveld fits to X-ray powder diffraction data¹⁷. p is the number of parameters adjusted in the fit, R_{WP} is the weighted profile crystallographic R factor, the F ratio is defined in the text, and F_{95} is the F ratio which would indicate a 95% ($\pm 2\sigma$) confidence limit on the improvement of the fit. For all Rietveld fits, we varied several diffraction lineshape parameters, the lattice parameters, the carbon isotropic thermal parameter, and anisotropic thermal parameters and fractional occupancy of the cation. The extra parameters for each KC₆₀ fit are the intensity, lattice parameters, and lineshape parameters of coexisting C₆₀ and K₃C₆₀.

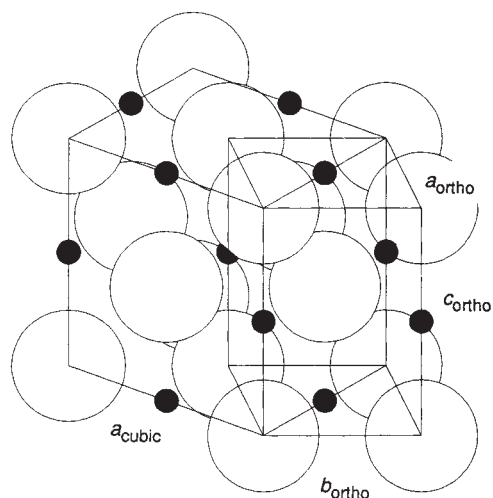


FIG. 1 Relationship between face-centred-cubic and body-centred-orthorhombic lattices. Fullerenes are indicated by large spheres, cations by smaller solid spheres. a_{cubic} is the lattice translation vector of the cubic phase; a_{ortho} , b_{ortho} and c_{ortho} those of the orthorhombic phase.

Samples were prepared by solid-state reaction of stoichiometric quantities of alkali metal and sublimed high-purity C_{60} powder. The sealed quartz capillaries allowed electron spin resonance (ESR) and X-ray diffraction measurements to be performed on the same samples. In particular, both samples show the same narrow ESR line described in ref. 4 for the Rb phase. X-ray powder diffraction spectra (Fig. 2) were collected at beamline X3B1 of the National Synchrotron Light Source, Brookhaven National Laboratory. Results from the two samples are very similar; we will concentrate our discussion on RbC_{60} .

Typically, in single-crystal structure determinations, the number of observed intensities vastly outweighs the number of structural parameters. The position of each atom in a refined structure is therefore determined more-or-less independently of the others, and the estimated standard error in the position of each atom is a fair representation of the uncertainty in its location. In contrast, if one tries to refine all the atomic positions in the X-ray powder diffraction data sets presented here, the calculation is unstable. We therefore proceed through a series of incremental improvements to the structural model, and outline these steps

in succession to illustrate the degree of confidence in each step. The results are summarized in Table 1.

The largest set of symmetry elements of icosahedral C_{60} which can be accommodated in an orthorhombic lattice is three mirror planes, if the C_{60} is located with three 2-fold axes along the coordinate directions in the space group $Immm$. The quality of this fit (Table 1, model 1 and ref. 4) is adequate to establish that the fullerenes and cations are roughly in the right positions, but it is clearly unacceptable as a structural determination. We experimented next with rotating the same undistorted (single bonds, 1.45 Å; double bonds, 1.39 Å) C_{60} in various directions. The results clearly favour a model in which one molecular 2-fold axis is aligned with the 9.13 Å a axis, and the perpendicular 2-fold axes are rotated by $45 \pm 5^\circ$ about a from the other two crystallographic axes (model 2). Other orientations, such as staggered pentagon-pentagon contacts along the a axis, give significantly worse fits.

The most plausible orthorhombic space group to describe a pseudo-body-centred array of rotated fullerenes is

$$P2_1/m2_1/n2_1/n,$$

with the fullerenes at $(0, 0, 0)$ and $(1/2, 1/2, 1/2)$ rotated in opposite directions. For an undistorted fullerene in the lattice, the nearest interfullerene carbon-carbon contact is 2.23 Å. This is too long for any C-C bond, but much shorter than the 3.3 Å van der Waals diameter characteristic of graphite or carbon atoms in aromatic molecules. We therefore relaxed the position of this frontier atom (C1 in Fig. 3), restricting it first to move along the a axis. It moved 0.37 Å towards the neighbouring fullerene (model 3), indicating a direct chemical bonding (C-C distance 1.49 Å) between fullerenes. Allowing this atom next to move freely in the molecular mirror plane indicated in Fig. 3, that is, rotated 45° from the crystal b - c plane, produces a slight further improvement (model 4) as the intra-fullerene C-C distance increases to 1.89 Å. For comparison, we relaxed the other 15 carbon atoms, one at a time, and found that none of them gave close to the same improvement in the fit. It is natural to consider next the propagation of that distortion into the fullerene, by relaxing atoms C2 and C3 (model 5). Here, the improvement in the quality of fit is less dramatic, but still noticeable. Atomic coordinates and selected bond lengths for this model are given in Table 2.

It is a truism that the more structural parameters that are refined, the better will be the agreement with an experimental profile. In order to gauge the improvement observed when one increases the number of refined parameters from q to p , obtaining a decrease of the weighted profile R factor from R_q to R_p ,

FIG. 2 X-ray diffraction pattern of RbC_{60} . X-rays of wavelength 1.14871 Å were selected by a double Si(111) monochromator, and the diffracted beam analysed by reflection from a Ge(111) crystal before a NaI scintillation detector. This Rietveld fit¹⁷ has a weighted profile R factor of 4.85%, unweighted $R = 3.80\%$, Bragg $R = 2.23\%$, $\chi^2 = 4.5$.

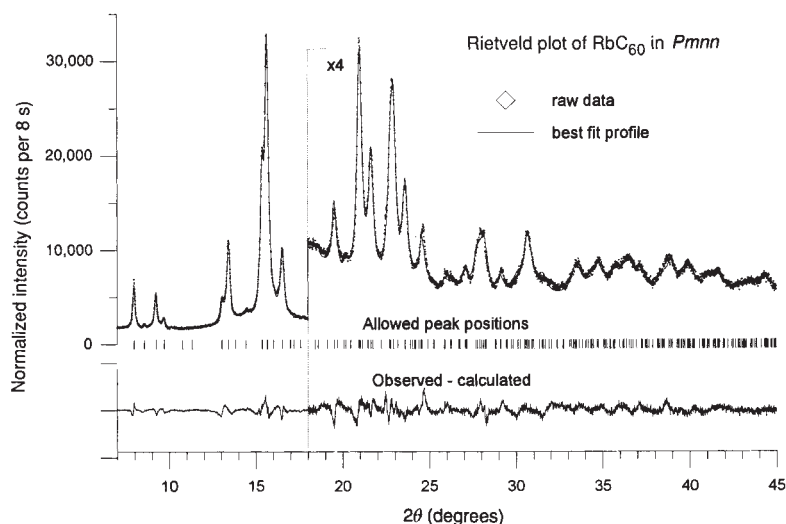


TABLE 2 Structural data derived from Rietveld fits

RbC₆₀	
Lattice Parameters (Å)	9.138 (5) × 10.107 (5) × 14.233 (5)
Space group	<i>P</i> 2/m 2 ₁ /n 2 ₁ /n (non-standard setting of no. 58)
Occupancy	Rb: 0.91 ± 0.04
Atomic coordinates	Rb: 0 0 1/2
	C1: 0.4214 0.0671 0.0467
	C2: 0.3410 0.0067 0.1293
	C3: 0.3410 0.1822 0.0023
	C4: 0.2833 -0.1082 0.1508
	C5: 0.2833 0.2102 -0.0797
	C6: 0.2514 0.1328 0.1636
	C7: 0.2514 0.2329 0.0912
	C8: 0.1554 -0.1267 0.2090
	C9: 0.1554 0.2918 -0.0939
	C10: 0.1279 0.1149 0.2198
	C11: 0.1279 0.3116 0.0774
	C12: 0.0791 -0.0172 0.2429
	C13: 0.0791 0.3416 -0.0168
	C14: 0.0763 -0.2402 0.1738
	C15: 0 0.1965 0.2056
	C16: 0 0.2931 0.1356
C1-C1 inter-fullerene (Å)	1.44 ± 0.15 (standard deviation)*
C1-C1 intra-fullerene (Å)	1.90 ± 0.15*
C1-C2, C1-C3 (Å)	1.51 ± 0.15*
Grain size (Å)	>500
Strain	~3% FWHM
Rb thermal ellipsoid (Å r.m.s.)	0.5 × 0.1 × 0.3
C thermal parameter	<i>B</i> = 2 ± 1
KC₆₀	
Lattice Parameters (Å)	9.109 (5) × 9.953 (5) × 14.321 (5)
K occupancy	1.00 ± 0.04
Impurity phases (mole fraction)	C ₆₀ , 26% K ₃ C ₆₀ , 10%
C1-C1 inter-fullerene (Å)	1.65 ± 0.20*
C1-C1 intra-fullerene (Å)	1.74 ± 0.20*
C1-C2, C1-C3 (Å)	1.50 ± 0.20*

* Error limits on bond lengths were obtained from fits in which the intra- or inter-fullerene C1-C1 bond length was fixed while the remaining 19 parameters were varied. The results of the *F*-test show that the degradation of this fit is statistically significant at the 95% (2σ) confidence level if *R*_{wp} increases from 4.85% to 5.02% for one fewer parameter.

we calculate the *F* statistic, defined as

$$F = (n-p)(p-q)^{-1} [(R_q/R_p)^2 - 1],$$

where *n* is the number of statistically independent observations⁹. The addition of new structural parameters is regarded as significant if the cumulative distribution function $\Psi(F, n-p, p-q)$ exceeds some value, say 95%. In single-crystal refinements, this is known as Hamilton's criterion. Its direct application to powder refinements is problematical, because the intensity at nearby points of a powder profile are not statistically independent. As a conservative estimate, we determine *n* by counting all Bragg peaks that are within one full-width at half-maximum as a single observation. In this way, the 3,801 points in the refined part of the profile of Fig. 2, covering 228 allowed Bragg peaks, represent *n* = 75 observations.

So far, this structure has been analysed without molecular disorder beyond a Debye-Waller factor to account for thermal vibrations. An alternative description is that each chain is randomly flipped about the *a* axis, so that the plane of the cyclobutane linkage is orientated either +45° or -45° from the *b* axis. Such binary orientational disorder is observed in other fullerenes¹⁰. In that case, the average structure would be described by the space group *Immm*, with a 50% occupation probability for each carbon. The quality of Rietveld fit and the

refined parameters are nearly identical (Table 1, model 6). From ¹³C NMR, it is known that the fullerenes are not flipping rapidly at room temperature, implying that any such disorder is static¹¹.

We have performed an equivalent analysis on the KC₆₀ sample. The results are summarized in Tables 1 and 2. Although there are minor differences between the two structures, the molecular bonding geometry is consistent for the two cations.

The interfullerene C1-C1 distance of 1.44 ± 0.15 Å (standard deviation) is within the usual range for C-C bonds, whereas the intra-fullerene distance of 1.90 ± 0.15 Å is significantly larger. Some caution is warranted in accepting such an unusual bond distance from X-ray powder diffraction data, but the result has been confirmed in several different samples. This may be compared with a recent single-crystal X-ray structure of dibromophenyl-methanofullerene, in which a single carbon atom bridges a pair of atoms on the fullerene, giving a distance between the bridgehead carbons of 1.76 ± 0.05 Å (ref. 12). In the present case, this anomaly suggests that a partial π-bond is formed between the cages at the expense of the intra-cage σ-bonds. The character of the inter-fullerene bond can be explained by the possible bonding structures of the four-membered ring shown at the bottom of Fig. 3. In **1** there is no inter-cage bond, although the C₆₀ molecules must be strained. Structure **2** is a symmetric (square) ring with four σ-bonds. In structure **3**, the intra-fullerene C1-C1 bond is broken, and a π-bond is formed between the fullerenes. Whereas **2** is predicted to have the lowest energy^{6,7} the other two structures may have contributions. A similar effect is observed in the strained bonds of some cyclophanes¹³. We note that short

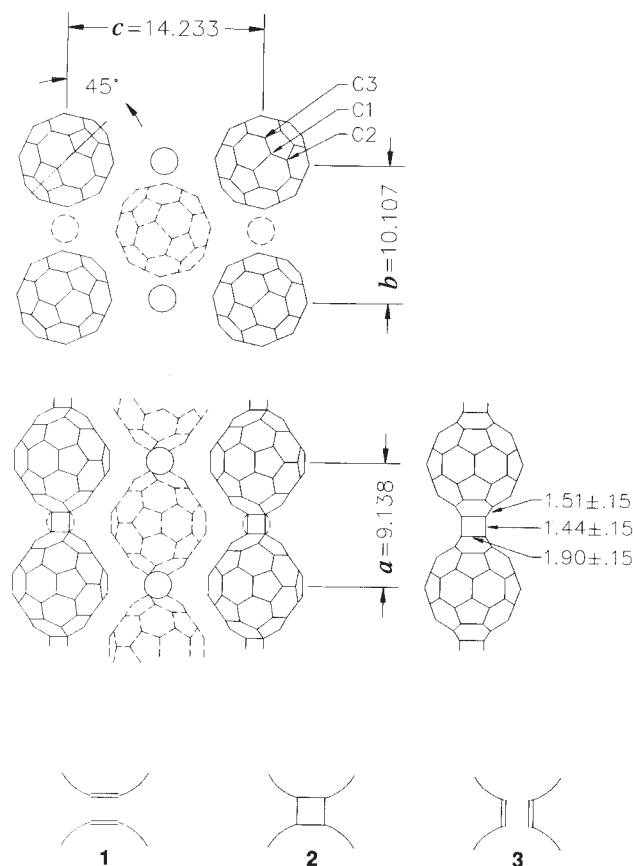


FIG. 3 Structure of RbC₆₀, showing covalently bonded chains along the *a* direction. Solid (broken) lines show fullerenes and cations centred at *x* = 0 (*x* = 1/2) in the *b*-*c* plane at the top, and *y* = 0 (*y* = 1/2) in the *a*-*c* plane in the middle. Solid line through the upper-left fullerene denotes the molecular mirror plane, and the fragment to the right is a projection of one chain perpendicular to that plane. (All distances given in Å.) Bottom, sketches showing three possible bonding structures of a four-membered ring.

inter-fullerene distances of 9.62 and 9.22 Å, suggestive of inter-molecule bonds, have also been reported for two pressure-synthesized phases of pure C₆₀ (ref. 14).

One puzzle in the interpretation of the structures of orthorhombic AC₆₀ as covalently bonded chains of fullerenes has been the low stability of this phase. As noted above, they depolymerize into fullerides with the rock-salt structure at 350–450 K, and the heat of formation, 12 J g⁻¹ (ref. 3) is only twice the heat of transformation at the 250 K orientational ordering transition of van der Waals C₆₀ (ref. 15). The contribution of the strained structure **3** may explain this low stability. Photopolymerized C₆₀ films also revert to the monomers at 400–450 K (ref. 16), bolstering the interpretation that the inter-fullerene bonding is similar in that system. Furthermore, because of the partial π -character of the inter-cage bonds, the conjugation of π -electrons extends through the polymer chains, supporting the conclusion that this material is a one-dimensional metal⁴. □

Received 9 June; accepted 26 July 1994.

1. Murphy, D. W. et al. *J. Phys. Chem. Solids* **53**, 1321–1332 (1992).
2. Zhou, O. & Cox, D. E. *J. Phys. Chem. Solids* **53**, 1373–1390 (1992).
3. Zhu, Q. et al. *Phys. Rev. B* **47**, 13948–13951 (1993).
4. Chauvet, O. et al. *Phys. Rev. Lett.* **72**, 2721–2724 (1994).
5. Rao, A. M. et al. *Science* **259**, 955–957 (1993).
6. Pekker, S., Forró, L., Mihály, L. & Jánosy, A. *Solid St. Commun.* **90**, 349–352 (1994).
7. Matsuzawa, N., Ata, M., Dixon, D. A. & Fitzgerald, G. J. *Phys. Chem.* **98**, 2555–2563 (1994).
8. Menon, M., Subbaswamy, K. R. & Sawtari, M. *Phys. Rev. B* **49**, 13966–13969 (1994).
9. Prince, E. & Spiegelman, C. H. in *International Tables for Crystallography Vol. C* (ed. Wilson, A. J. C.) 618–624 (Kluwer Academic, Dordrecht, 1992).
10. Stephens, P. W. et al. *Nature* **351**, 632–634 (1991).
11. Tycko, R., Dabaghi, G., Murphy, D. W., Zhu, Q. & Fischer, J. E. *Phys. Rev. B* **48**, 9097–9105 (1993).
12. Iverson, B. et al. *J. Am. chem. Soc.* (submitted).
13. Keehn, P. N. in *Cyclophanes* (eds Keehn, P. M. & Rosenfeld, S. M.) Ch. 3 (Academic, New York, 1983).
14. Iwasa, Y. et al. *Science* **264**, 1570–1572 (1994).
15. Heiney, P. A. et al. *Phys. Rev. Lett.* **66**, 2911–2914 (1991).
16. Wang, Y., Holden, J. M., Bi, X. & Eklund, P. C. *Chem. Phys. Lett.* **217**, 413–417 (1994).
17. Larsen, A. C. & Von Dreele, R. B. Report No. LAUR 86-748 (Los Alamos National Lab., 1990).

ACKNOWLEDGEMENTS. We thank L. Mihály, P. Coppens, J. Kaduk, K. Gardner, R. Harlow, A. Larson and R. Von Dreele for discussions and assistance. This work was supported by the US National Science Foundation, the Hungarian National Science Foundation, the US-Hungarian Joint Fund and the Swiss National Foundation of Scientific Research. The SUNY beamline at the National Synchrotron Light Source is supported by the US Department of Energy.

Global decrease in atmospheric carbon monoxide concentration

M. A. K. Khalil & R. A. Rasmussen

Global Change Research Center and Department of Environmental Science and Engineering, Oregon Graduate Institute, Beaverton, Oregon 97006, USA

CARBON monoxide plays an important role in the oxidizing capacity of the Earth's atmosphere, and may thereby indirectly affect the concentrations of many man-made and natural trace gases, which in turn affect climate, atmospheric chemistry and the ozone layer¹. CO is produced in the atmosphere by the oxidation of methane and other hydrocarbons, and is released into the atmosphere from automobiles, agricultural waste and the burning of savanna^{1–4}. Recent estimates^{1–4} show that human activities such as these are presently responsible for more than half the annual emissions of CO. During the 1980s there was evidence that atmospheric CO concentrations were increasing at $\sim 1.2 \pm 0.6\%$ per year, leading to feedbacks that could amplify global warming. Here we present a continuation of these measurements which show that from 1988 to 1992 global CO concentrations have started to decline rapidly at a rate of about $-2.6 \pm 0.8\%$ per year. A recent

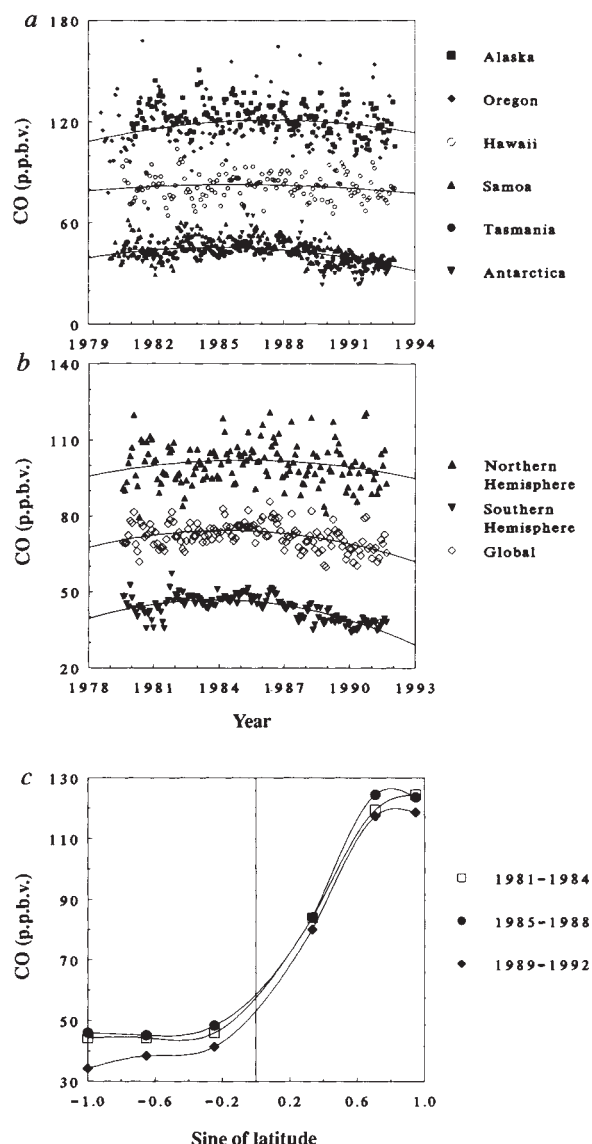


FIG. 1 Concentrations of CO at six sites representing polar, middle and tropical latitudes of both hemispheres. The sites are: Pt Barrow, Alaska; Cape Meares, Oregon; Cape Kumukahi and Mauna Loa Observatory, Hawaii; Cape Matatula, Samoa; Cape Grim, Tasmania; Palmer Station and South Pole, Antarctica. *b*, Hemispherical and global averages; *c*, latitudinal distribution. (This Figure is a continuation of data in ref. 6). The experimental conditions were as follows. Sampling: triplicate air samples collected at each site in 0.8-l electropolished stainless-steel containers and analysed at Oregon Graduate Institute by gas chromatography (Carle 211-MS GC/FID and Trace Analytical HgO GC/RGD). Measurements before 1984 were by GC/FID and afterwards by both instruments. Triplicates usually agree within 1–2%. If one is $\geq 20\%$ different from the other two, it is eliminated (very few were discarded). Concentrations are stable for >1 yr in sampling containers. Detection limits: 5 p.p.b.v. (RGD), 20 p.p.b.v. (FID). Linearity: 20 p.p.b.v. to tens of p.p.m.v. (FID), 5–200 p.p.b.v. for RGD (also approximately linear between 200–700 p.p.b.v.). Standards: original primary standards, all in synthetic air, from National Bureau of Standards (NBS), now NIST; Scott Marin, Matheson, Alpha Gas: Concentrations were 321 p.p.b.v. to 10 p.p.m.v. with certified accuracy $\pm 2\%$. Working Standards: ambient air in 32-l high-pressure stainless-steel tanks starting at 400–600 p.s.i.g. (pounds per square inch, gauge) and not used when below 200 p.s.i.g. All working standards were calibrated against the NBS standard reference material (SRM 1677B, obtained in 1984) at 10 p.p.m.v. and checked periodically against this standard and against each other. Precision 1% (measured once per month). Blanks: tests on contamination from lines and pumps using synthetic air give CO concentrations below detection limits of 5 p.p.b.v. Trends in standards: 1980–86, $-0.5 \pm 0.7\%$ yr⁻¹; 1987–93, $0.08 \pm 1.7\%$ yr⁻¹ and overall 1980–93, $0.2 \pm 0.3\%$ yr⁻¹.

study⁵ has verified our findings with data from the past 3–4 years. The rate of decrease is particularly rapid in the Southern Hemisphere; we hypothesize that this may reflect a reduction in tropical biomass burning. The total amount of carbon monoxide in the atmosphere is less now than a decade ago.

Measured CO concentrations are shown in Fig. 1 (refs 6, 7). Hemispherical values are calculated as area-weighted means of concentrations at the sites representing polar, middle and equatorial latitudes. Carbon monoxide increased, as previously reported⁷, until about 1987; then concentrations started to fall.

The trends (Fig. 2) are calculated by several complementary techniques⁷. We divided the data into two 6-year periods: January 1981 to January 1987 and January 1987 to January 1993 (for Cape Meares, we used the longer record starting at August 1979). First we subtracted the seasonal cycles from the data; then, to estimate trends, we used ordinary linear regression, non-parametric trend analysis and a difference method^{8,9}.

We find that in the first half of the record all methods give increasing trends at all locations, although at some sites the trends are small or statistically not significant at the 5% level. Nonetheless, positive trends at all widely dispersed locations are not likely to occur by chance (the probability is 1.6% assuming that a negative or positive trend is equally probable, if due entirely to chance). The increasing trend was largest at the northern mid-latitude site of Cape Meares in Oregon, which is in the latitude band with highest anthropogenic emissions^{3,4}. Large increases were also observed in the Antarctic data; however, these data are the fewest and least reliable, especially during the first 5 years of the record. In the second half of the data set, there is a consistent decrease of CO at all sites that is statistically significant everywhere and by all three methods (at the 5% level). The decrease during recent years is therefore much more definitive than the increasing trends were in the first part of the data. Because of the fast recent decreases, there is a net loss of CO in the atmosphere during the course of our 12-yr experiment, resulting in -4 parts per billion by volume (p.p.b.v.), or -5% less CO in the atmosphere during the last 3 years (1990–92) compared to the first 3 years (1981–82).

Our results on the recent decreases of CO, were verified⁵ by the NOAA climate monitoring and diagnostic laboratory (CMDL) using a shorter data set; their estimated rates of decrease however, are more rapid than ours. We further verified the trends at Cape Grim by comparing our data to independent measurements taken by CSIRO. For the period after CO started decreasing, published data are available from CSIRO between 1987–1991¹⁰. The decreasing trend (1987–91) in the CSIRO data is -1.8 ± 0.7 p.p.b.v. yr^{-1} and it is -2.4 ± 0.8 p.p.b.v. yr^{-1} in our data. The two data sets have relatively small disagreements during this period and for some periods before 1987 (when CO was not decreasing). Although these comparisons are encouraging, there are still many uncertainties in the estimate of the decrease but less so in the idea that CO is now decreasing and has been for the past 4–6 years.

The trends of CO, although extremely important in understanding global atmospheric chemistry, have been fairly small. The annual average concentration of CO has a mean of

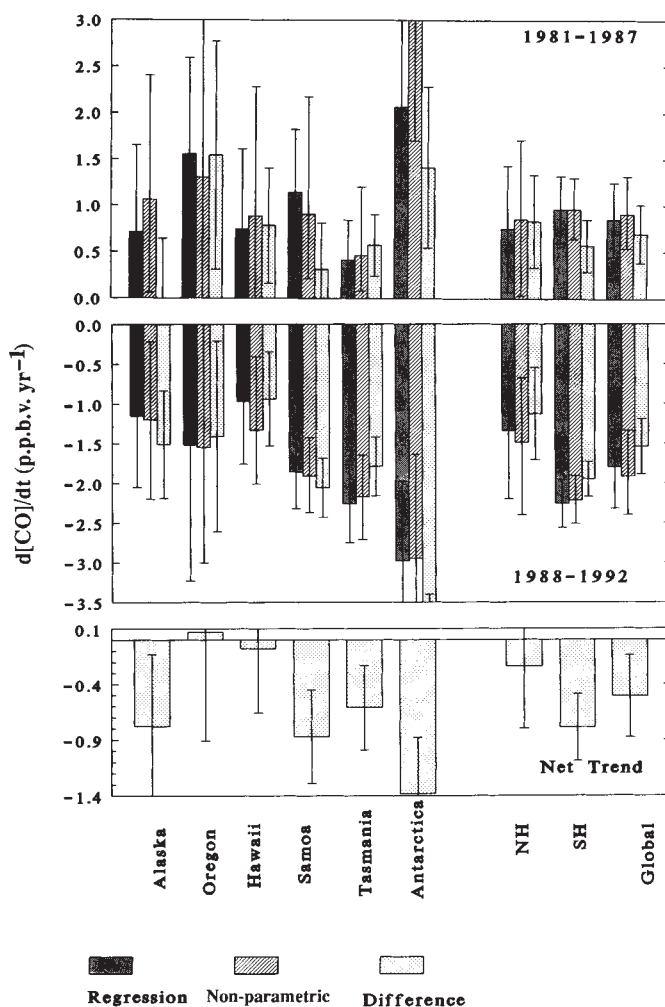


FIG. 2 Global trends of CO (p.p.b.v. yr^{-1}). Results are shown for three methods: regression, non-parametric and a difference method. In the difference method, we obtained average concentrations $C(i, j)$ during three 4-yr periods; 1981–84, 1985–88 and 1989–92 (years inclusive) where i is period 1 to 3 and j is site 1 to 6. The trend is $b(1, j) = [C(2, j) - C(1, j)] / \Delta T_1$ for the first period, $b(2, j) = [C(3, j) - C(2, j)] / \Delta T_2$ for the second period, and a net change $b(3, j) = [C(3, j) - C(1, j)] / [\Delta T_1 + \Delta T_2]$. The bottom panel shows the composite effect during the course of the experiment. Uncertainties in the trends are increased because of atmospheric, instrumental and standards variabilities. The long period of measurements compensates for this variability, leading to statistically significant estimates of trends.

74.7 p.p.b.v. and standard deviation of 3.5 p.p.b.v., resulting in a variability of less than $\pm 5\%$ (standard deviation divided by the mean during the course of the experiments. The difference between the maximum and minimum annual globally averaged concentrations observed during these 12 years is slightly less than 15%.

It will take more research to determine why the concentrations are declining rapidly in recent years. Here we discuss some possibilities. Because of its short atmospheric lifetime of 2–3 months, the concentration of CO is in approximate equilibrium with the production and destruction (a pseudo steady state). If either the production or destruction rates change, the concentration of CO responds quickly, retaining no memory of past trends of sources and sinks. This is unlike longer-lived gases, such as methane or the chlorofluorocarbons, that continue to build up in the atmosphere long after sources or sinks become constant or even after the sources start to decline. To sustain an increasing concentration of CO requires a constantly increasing source or a decreasing removal process. Also, because of its short lifetime,

TABLE 1 Estimated average trends of CO between 1981 and 1992

	1981–86		1987–92	
	b (% yr^{-1})	B (p.p.b.v. yr^{-1})	b (% yr^{-1})	B (p.p.b.v. yr^{-1})
NH	0.8 ± 0.7	0.8 ± 0.7	-1.4 ± 0.9	-1.3 ± 0.9
SH	2.1 ± 0.8	1.0 ± 0.4	-5.2 ± 0.7	-2.3 ± 0.3
Global	1.2 ± 0.6	0.9 ± 0.4	-2.6 ± 0.8	-1.8 ± 0.5

The models used with standard regression analyses are $C(\text{p.p.b.v.}) = A(\text{p.p.b.v.}) + B(\text{p.p.b.v. yr}^{-1}t(\text{yr}))$ and $C(\text{p.p.b.v.}) = A(\text{p.p.b.v.}) \exp[b(\% \text{yr}^{-1})/100 t(\text{yr})]$. Table shows $B(\text{p.p.b.v. yr}^{-1})$ and $b(\% \text{yr}^{-1})$. $\pm$ values are 90% confidence limits; NH, Northern Hemisphere; SH, Southern Hemisphere.

emissions in one region may not substantially affect the concentrations or trends far away, particularly across the hemispheres.

CO is removed mostly by reacting with atmospheric OH. Under the assumption of pseudo steady state for the global average concentration, $[CO] \approx S/K_{\text{eff}}[OH]$ where S represents the emissions, and K_{eff} is the effective average rate constant for the reaction of CO with OH. From this equation the trends of CO are $\% \delta[CO]/\delta t = \{\% \delta S/\delta t - \% \delta[OH]/\delta t\} / \{1 - \% \delta[OH]/\delta t\} \approx \% \delta S/\delta t - \% \delta[OH]/\delta t$ (where $\% \delta x/\delta t$ is the percentage change of x over time δt). The changes in emissions S and in OH concentration have an additive effect on the trend of CO. A decreasing concentration of CO occurs if the emissions decline or if OH concentrations rise.

Our analysis suggests that the observed changes are occurring mostly because of decreasing emissions, although the sources from which emissions are decreasing may be quite different between the two hemispheres. In the Northern Hemisphere most of the anthropogenic sources of CO have been decreasing in recent years. Because CO is a "criteria pollutant" in the United States, most urban sources such as automobiles and CO precursors have been under increasingly stringent controls. Controls have extended to Europe and elsewhere. The slowdown in regional industrial processes, such as in the former USSR may also be a factor in reducing CO concentrations. Even the oxidation of methane is contributing to smaller increases of CO than before, because methane increase has slowed down considerably¹¹. These factors taken together may be enough to account for the decrease of CO in the Northern Hemisphere.

The alternative explanation is the increase of OH. For some time it was thought that OH concentrations must be decreasing because of large and increasing emissions of methane and CO from human activities. But recently it has become apparent that there are compensating processes that boost OH production^{12,13}, and it has now been argued that OH may be increasing^{14,15}. One proposed mechanism is that as stratospheric ozone decreases, possibly from the effects of the chlorofluorocarbons, more solar radiation of wavelengths <330 nm reaches the troposphere, stimulating OH production. This may be particularly significant in the Antarctic and the deep southern Hemisphere because of the ozone hole. Measurements from the TOMS instrument on the Nimbus 7 satellite show declines of ozone mostly at middle and high latitudes of both hemispheres¹⁶. Furthermore, increases of tropospheric ozone and other precursors may be adding to increasing OH in some places. If OH is increasing in some parts of the atmosphere, or everywhere, it may contribute to the observed decreasing trend of CO under some conditions. In regions where more CO comes from direct emissions than the oxidation of methane, increasing OH would lead to a decrease of CO. But in regions where the dominant source of CO is oxidation of methane, which is still increasing in the atmosphere, oxidation of methane would lead to an increase of CO. Moreover, the decreases of CO at $1.5\% \text{ yr}^{-1}$ in the Northern Hemisphere and $>5\% \text{ yr}^{-1}$ in the Southern Hemisphere are too rapid compared to the estimated increase of OH from stratospheric ozone depletion.

The rapid decline observed in the middle and higher latitudes of the Southern Hemisphere, represented by measurements taken at Tasmania and Antarctica, is puzzling at first. The tropics, because of high OH concentrations, act as a trap for the transport of anthropogenic CO from the middle and high northern latitudes to the Southern Hemisphere as evidenced by the sharp reduction of CO concentrations across the equator (Fig. 1). The decrease of CO in the middle and high southern latitudes, therefore, cannot be explained by decreasing emissions in the Northern Hemisphere or the action of increased OH there. Increases of OH in this region also are not sufficient to explain the decrease of CO because there are hardly any direct sources and the greater removal of CO by increasing OH is compensated by the greater production from the oxidation of methane. This leaves only the effect of transport from the southern tropical

and sub-tropical latitudes. Calculations show that the decreasing trend of CO at Tasmania is explained mostly by transport of CO from the tropics represented by measurements at Samoa. Many details of the changing concentrations of CO at Tasmania are reflected also in the measurements at Samoa leading to a relatively high correlation of ~ 0.5 (after subtracting seasonal cycles). We think that a decreasing trend of tropical biomass burning, which is the dominant source of CO in the tropics¹, is a possible cause of the decreasing CO in the middle and deeper Southern Hemisphere. Since the decrease of emissions in the southern tropics leads to the decrease of concentrations at higher southern latitudes, the decrease of CO in the entire Southern Hemisphere may be driven by decreases in tropical biomass burning. The latitudinal details of the decreasing trend are complex, and the role of biomass burning in causing these changes awaits further verification. Whether the decreases of biomass burning during the past 6 years, if they are occurring, are a permanent or temporary situation or whether they are caused by cycles of climate or changes in human activities, we cannot say from the data at hand. What is clear is that the rather sudden and rapid decline in CO is an indicator of global change with far-reaching implications regarding the effect of human activities on the global environment. □

Received 18 August 1993; accepted 20 July 1994.

1. Logan, J. A., Prather, M. J., Wofsy, S. C. & McElroy, M. B. *J. geophys. Res.* **86**, 7210–7254 (1981).
2. *Atmospheric Ozone 1985* (World Meteorological Organization, Geneva, 1986).
3. Khalil, M. A. K. & Rasmussen, R. A. *Chemosphere* **20**, 207–242 (1990).
4. *Air Quality Criteria for Carbon Monoxide* (US Environmental Protection Agency, Washington DC, 1990).
5. Novelli, P. et al. *Science* **263**, 1587–1590 (1994).
6. Rasmussen, R. A. & Khalil, M. A. K. *J. geophys. Res.* **86**, 9826–9832 (1981); in *Proc. NATO Advanced Study Institute on Atmospheric Ozone* (ed. Aikin, A. E.) 209–252 (Department of Transportation, Washington DC, 1980).
7. Khalil, M. A. K. & Rasmussen, R. A. *Nature* **332**, 242–245 (1988).
8. Hollander, M. & Wolfe, D. *Nonparametric Statistical Methods* (Wiley, New York, 1973).
9. Snedecor, G. W. & Cochran, W. G. *Statistical Methods* (Iowa State Univ. Press, Ames, 1980).
10. Fraser, P., Coram, S. & Derek, N. *Baseline Atmospheric Program 1990* (Div. of Atmospheric Research CSIRO, Melbourne, 1992).
11. Khalil, M. A. K. & Rasmussen, R. A. *Envir. Sci. & Technol.* **24**, 549–553 (1990); *Chemosphere* **26**, 803–814 (1993).
12. Pinto, J. P. & Khalil, M. A. K. *Tellus* **43B**, 347–352 (1991).
13. Thompson, A. M. *Science* **256**, 1157–1165 (1992).
14. Madronich, S. & Granier, C. *Geophys. Res. Lett.* **19**, 465–467 (1992).
15. Prinn, R. G. et al. *J. geophys. Res.* **97**, 2445–2461 (1992).
16. *Scientific Assessment of Ozone Depletion: 1991* 2.23 (World Meteorological Organization, Geneva, 1991).

ACKNOWLEDGEMENTS. We thank the NOAA/CMDL and Cape Grim/CSIRO staff for collecting samples at most of the sites. We thank D. Stearns for laboratory work, and M. Shearer and F. Moraes for discussions. This work was supported by the US NSF and the US Department of Energy, with additional support from the BRC and Andaz Co.

Transitions between explosive and effusive eruptions of silicic magmas

Andrew W. Woods* and Takehiro Koyaguchi†

* Institute of Theoretical Geophysics, 20 Silver Street, University of Cambridge, Cambridge CB3 9EW, UK

† Earthquake Research Institute, University of Tokyo, Yayoi, Tokyo 113, Japan

WATER-RICH silicic magmas may erupt explosively, giving rise to massive columns of fragmented ash^{1–5}, or effusively as viscous, bubbly lavas to form lava domes^{6–8}. Complex cycles of explosive and effusive eruptions are often observed^{8–12}. Although these two eruption mechanisms have been modelled independently^{1,8}, the conditions under which a volcano exhibits a particular style of eruption are not known and the transitions between eruption styles are poorly understood. By modelling the ascent of magma along a permeable conduit, we show here that, for small magma-chamber overpressures, the style of eruption—explosive or effusive—is

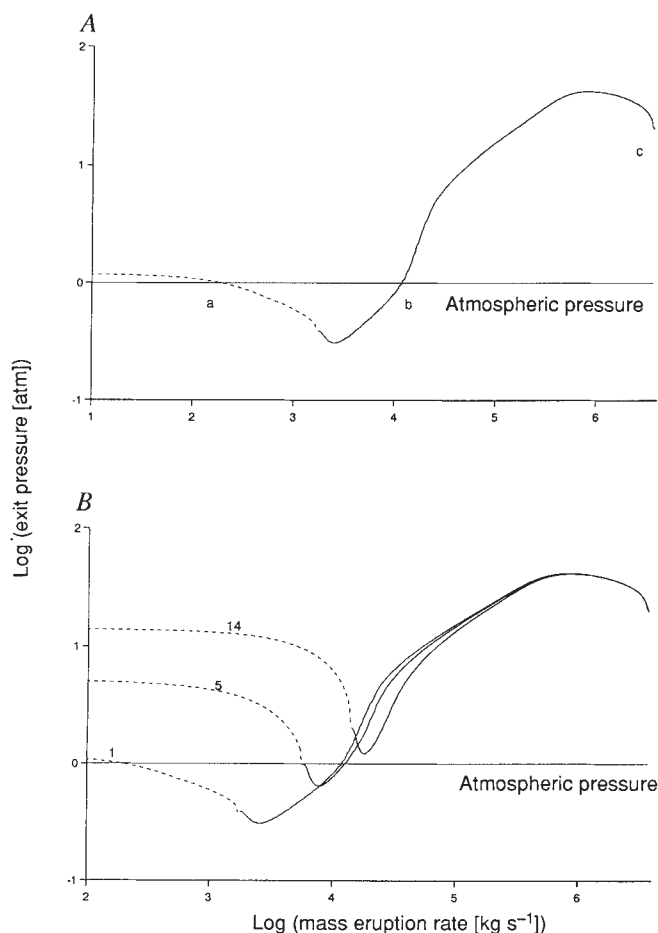


FIG. 1. Calculation of the exit pressure as a function of the mass eruption rate. A, chamber overpressure is 1.0 atm; B, chamber overpressures of 1.0, 5.0 and 14.0 atm. The dashed part of each curve corresponds to eruption of a vesicular foamy lava; the solid part to a fragmented magma volatile mixture. In these calculations, the conduit radius is 10 m, the chamber depth is 2 km, the initial volatile content of the magma is 3%, the permeability of the conduit walls is $10^{-12.5} \exp(-z/1,000) \text{ m}^2$ where z is the distance below the surface, in metres. The magma and country rock densities are set to $2,500 \text{ kg m}^{-3}$; the overpressure is the pressure in excess of lithostatic. Solutions a and b correspond to effusive venting at atmospheric pressure. Solution c corresponds to explosive venting above atmospheric pressure.

dependent on the magma flow rate. We also identify a critical overpressure above which only explosive eruptions occur. Complex eruption sequences follow naturally. For example, if a shallow magmatic system is supplied with magma at an intermediate flow rate, the eruption will be slow and effusive until the chamber overpressure becomes too large. An explosive eruption then relieves the overpressure and an effusive eruption style resumes. We suggest that monitoring of temporal variations in chamber overpressure (for example, by measuring ground deformation) can be used to assess whether a passively effusing volcano has the potential to erupt explosively.

As water-rich silicic magma ascends to the surface and decompresses, volatiles are exsolved. If the conduit walls are impermeable, the volatiles remain trapped in the magma which inflates and eventually fragments. Wilson *et al.*¹ showed that this leads to the explosive eruption of a sonic, or even supersonic dusty jet, at a rate of 10^6 – 10^9 kg s^{-1} and a velocity of 100–200 m s^{-1} . In contrast, Jaupart and Allegre⁸ showed that if the conduit walls are permeable, then for very small (effusive) eruption rates, 0.1 – 10^3 kg s^{-1} , many of the volatiles can escape and a foamy liquid or possibly fragmented magma effuses from the

vent at speeds of order 10^{-5} – 10^{-2} m s^{-1} . However, each of these results suggests that there is a unique eruption rate for given conditions in a magma chamber, and they do not explain or predict whether a volcano will erupt explosively or effusively. By further examining the ascent of magma along a permeable conduit, we now address these fundamental issues.

As in ref. 8, we assume that the rate of loss of magmatic volatiles from the magma to the permeable conduit walls, Q_w , equals the rate of displacement of liquid through the associated permeable rock per unit depth

$$Q_w = 2\pi r p_w K (p - p_h) / \mu L \quad (1)$$

where r is the conduit radius, K the permeability, p the magma pressure, p_h the far-field pressure in the permeable rock (assumed to be hydrostatic¹³), μ_w and ρ_w the viscosity and density of water, and L is a typical length scale for the flow in the permeable rock. Owing to the paucity of data, for simplicity, we assume the permeability decays exponentially with depth, with an e-folding scale of 1 km (refs 8, 9, 14). The conservation of mass flux, Q , and gas flux, nQ , in the ascending magma–volatile mixture then has the form

$$\frac{dQ}{dz} = -Q_w; \quad \frac{dnQ}{dz} = -Q_w - Q_m \frac{dn_m}{dz} \quad (2, 3)$$

where Q_m is the magma flux, n is the gas mass fraction and n_m is the mass fraction of volatiles dissolved in the magma, as given by Henry's law, $n_m = sp^\beta$. For H_2O , $s = 4.1 \times 10^{-6} \text{ kg}^{0.5} \text{ m}^{0.5} \text{ s}^{-1}$ and $\beta = 0.5$. The mass flux $Q = \pi r^2 \rho u$, where u is the upward speed and the ρ is the bulk density of the volatile–magma mixture^{1,5,8}. The model is completed by imposing conservation of momentum in the magma, which has the approximate form⁸

$$\rho u \frac{du}{dz} = -\frac{dp}{dz} - \rho g - fu$$

where fu is the frictional force and $g = 9.81 \text{ m s}^{-2}$. The parametrization of the friction factor f is described by Jaupart and Allegre⁸, and depends on whether the magma–volatile mixture is fragmented; fragmentation is assumed to occur when the volatile volume fraction exceeds 77% (ref. 15).

Flow solutions are found by numerical shooting¹: for a given magma-chamber pressure, the model predicts a variation of exit pressure with flow rate (Fig. 1A) but only those flows which exit at atmospheric pressure ('a' and 'b') or with the speed of sound ('c') are possible^{1,3}. The flow solution 'c', for which the material erupts at the speed of sound, corresponds to an explosive eruption^{1–3}. Owing to the high speed of the flow, only a very small fraction of the volatiles are driven into the surrounding permeable rock by the excess pressure in the conduit (equation (1); Fig. 2A). The conduit therefore behaves as if impermeable^{1,3}, and the magma inflates, fragments and accelerates to very high speed. The flow solution 'a' is analogous to the solutions presented by Jaupart and Allegre⁸. These correspond to a slow effusive eruption in which most of the exsolved volatiles escape into the conduit walls (Fig. 2) and the pressure remains close to lithostatic throughout the conduit⁸. The magma only becomes significantly inflated near the vent, where the pressures are low and the rate of volatile expansion is greatest. In the faster flow solution 'b', a smaller fraction of the volatiles escape from the conduit, so the magma inflates more, and may even fragment (Fig. 2). But because many of the volatiles do escape from the conduit, the flow 'b' is also weak, with eruption velocities of only 0.1 m s^{-1} . We class this as an effusive eruption (type 'b')⁷. In flow 'b', just below the vent the rate of escape of volatiles through the conduit walls increases above the rate of volatile production through magma decompression (equation (1)); this causes the void fraction to decrease and so the mixture density increases and velocity decreases (Fig. 2). The effusive solutions 'a' and 'b' are quite distinct from the

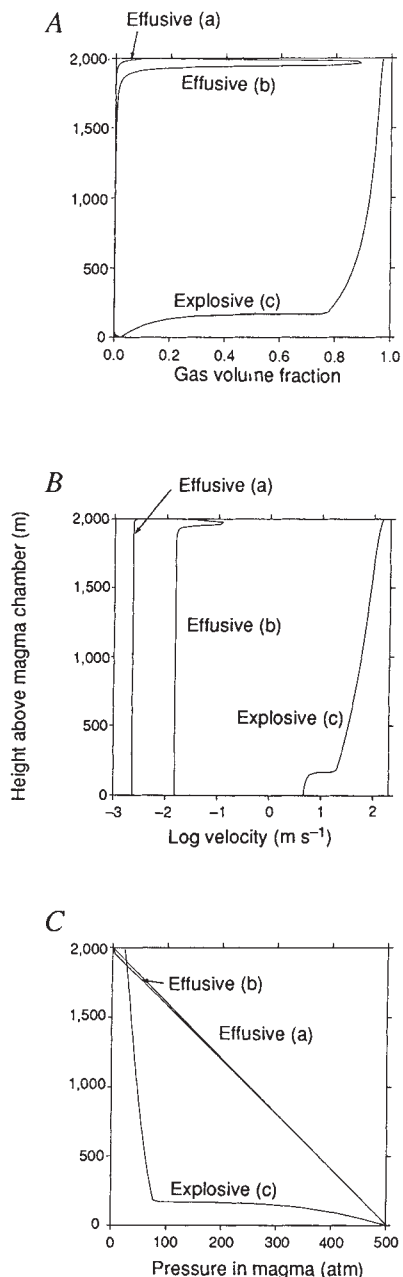


FIG. 2 Variation of A, void fraction of the magma-volatile mixture; B, velocity of magma-volatile mixture; C pressure of the magma-volatile mixture with the height above the magma chamber. In the explosive eruption, the pressure falls below lithostatic owing to the large frictional stresses below the fragmentation surface¹⁻³. In the effusive eruption 'b' the magma becomes more inflated than in solution 'a', and the velocity of the magma increases. In turn, there is a decrease in the density and hence work done against gravity which offsets the increased frictional force associated with the higher flow rate; thus the material can vent at atmospheric pressure in both solutions 'a' and 'b'.

explosive eruption 'c': the non-uniqueness emerges from the nonlinearity of the governing equations (1)–(4).

As the chamber pressure increases, the effusive flow rate 'a' increases towards the effusive flow rate 'b' (Figs 1A, 3). Although the higher speed increases the frictional pressure loss, fewer volatiles escape from the ascending magma. Thus the density of the mixture and hence the work done against gravity decrease, so the material can still issue at atmospheric pressure. Eventually, if the overpressure increases sufficiently, then the increase in the frictional pressure loss cannot compensate for

the decrease in the work done against gravity combined with the larger chamber overpressure; as a result, the magma-volatile mixture is predicted to vent with a pressure above atmospheric for all flow rates (overpressure 14 atm, Fig. 1B). Therefore, the effusive solutions 'a' and 'b' cease to be possible and the only style of eruption is explosive ('c') with the material issuing at the speed of sound and a pressure greater than atmospheric¹⁻³. The maximum overpressure for which effusive eruption is possible decreases with magmatic gas content (curves 1 and 2 in Fig. 3), but increases with conduit permeability as more volatiles are able to escape from the conduit for a given flow rate (curves 1 and 3 in Fig. 3).

The non-uniqueness of eruption rate for small chamber overpressures leads to dramatic transitions between effusive and explosive eruptions. For example, during some eruptions a deep magma source may feed a shallow magmatic system (Fig. 4a)^{7,9,16-18}. Suppose that magma is supplied to a magma chamber at a rate $Q_i(t)$, whereas it effuses to the surface at a slower rate $Q(\Delta P)$, which depends upon the overpressure of the chamber ΔP (Fig. 3), then magma accumulates in the chamber at a rate $Q_i - Q$. This increases the chamber overpressure ΔP according to^{19,20}

$$\frac{d\Delta P}{dt} = \mu(Q_i - Q)/\sigma V \quad (5)$$

where μ is the effective bulk modulus of the host rock, σ is the magma density, and V is the volume of the shallow reservoir. Eventually the overpressure builds up to the maximum value for which effusive eruption is possible; at this point an explosive eruption will develop. This relieves the overpressure in the shallow magmatic system, and the effusive phase of the eruption may then recommence (Fig 4b, c). If the rate of supply of magma, Q_i , gradually decreases, then the interval between such explosive cycles increases. Finally, when the supply rate has decayed to values comparable with the effusive eruption rate, the activity may end with a steadily waning effusive eruption (Fig. 4). For the typical parameters of Fig. 4, the explosive events erupt $\sim 10^6$ – 10^7 m³ of material over an interval of 10 – 10^3 s, whereas the effusive phase issues $\sim 10^5$ – 10^6 m³ of material over an interval of 10^5 – 10^6 s. Although this is an extremely simple model, and caution should be exercised interpreting the qualitative predictions, these numbers are consistent with observations of Mt St Helens during 1980^{11,12}.

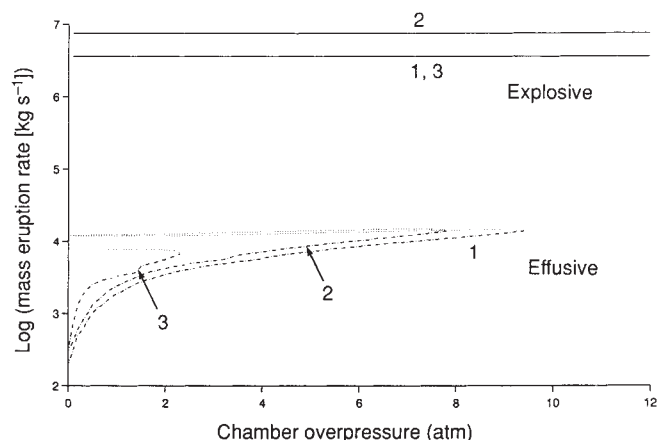


FIG. 3 Variation of the explosive (solid) and effusive (dashed-type 'a'; dotted-type 'b') mass eruption rates with the overpressure. Curves (1) correspond to a chamber 2 km below the surface, with conduit radius of 10 m and typical surface permeability^{8,14} of $10^{-12.5}$ m². The magma has an initial volatile content of 3%. In (2), the initial magmatic volatile content has been changed to 5% while in curves (3) the conduit permeability has value $10^{-12.75}$ m². The explosive eruptions in cases (1) and (3) are indistinguishable.

Under a lava dome, magma may vent effusively at higher pressures, of the order of 1–30 atm (refs 8, 20), but again our model predicts a maximum chamber overpressure for effusion. Therefore, if a lava dome collapses or explodes phreatically²¹, then the chamber pressure may be so high that after collapse, the only possible eruptive style is explosive. This may account for the powerful pyroclastic flows during the 1902 eruption of

Mt Pelee^{8,21}. Finally, noting the possible complications due to dome collapse, our model suggests that measurements of the temporal changes in chamber overpressure (based, for example, upon ground deformation) may be useful in assessing whether an effusing volcano will become explosive. □

Received 25 April; accepted 14 July 1994.

1. Wilson, L., Sparks, R. S. J. & Walker, G. P. L. *Geophys. J. R. astr. Soc.* **63**, 117–148 (1980).
2. Giberti, G. & Wilson, L. *Bull. volcan.* **52**, 515–521 (1987).
3. Dobran, F. J. *Volcan. geotherm. Res.* **49**, 285–311 (1992).
4. Wilson, L., Sparks, R. S. J., Huang, T. C. & Watkins, N. D. *J. geophys. Res.* **83**, 1829–1836 (1978).
5. Woods, A. W. *Bull. volcan.* **50**, 69–91 (1988).
6. Eichelberger, J. C., Carrigan, C. R., Westrich, H. R. & Price, R. H. *Nature* **323**, 598–602 (1986).
7. Newhall, C. & Melson, W. J. *Volcan. geotherm. Res.* **17**, 111–131 (1983).
8. Jaupart, C. & Allegre, C. J. *Earth planet. Sci. Lett.* **102**, 413–429 (1991).
9. Bursik, M. I. J. *Volcan. geotherm. Res.* **57**, 57–70 (1993).
10. Scandone, R. & Malone, S. D. J. *Volcan. Geotherm. Res.* **23**, 239–262 (1984).
11. Moore, J. G., Lipman, P., Swanson, D. & Alpha, T. R. *US Geol. Surv. Prof. Pap.* **1250**, 541–548 (1981).
12. Christianson, R. L. & Peterson, D. W. *US Geol. Surv. Prof. Pap.* **1250**, 17–30 (1981).
13. Attewell, P. & Farmer, I. *Principles of Engineering Geology* (Chapman & Hall, New York, 1976).
14. Heiken, G., Wohletz, K. & Eichelberger, J. C. J. *geophys. Res.* **93**, 4335–4350 (1988).
15. Sparks, R. S. J. *J. Volcan. geotherm. Res.* **3**, 1–17 (1978).
16. Wadge, G. J. *Volcan. geotherm. Res.* **11**, 139–168 (1981).
17. Barker, S. E. & Malone, S. D. J. *geophys. Res.* **96**, 11883–11894, (1991).
18. Blake, S. *Nature* **289**, 783–785 (1981).
19. Tait, S. R., Jaupart, C. & Vergnolle, S. *Earth planet. Sci. Lett.* **92**, 107–123 (1989).
20. Stasui, M., Jaupart, C. & Sparks, R. S. J. *Earth planet. Sci. Lett.* **114**, 505–516 (1993).
21. Fisher, R. V., Smith, A. L. & Robool, M. J. *Geology* **8**, 472–476 (1980).

ACKNOWLEDGEMENTS. A.W.W. was supported by the Japanese Education and Cultural Ministry and the British Council while visiting ERI, University of Tokyo. M. I. Bursik kindly suggested some useful references.

Perceptual sensitivity maps within globally defined visual shapes

Ilona Kovács & Bela Julesz

Laboratory of Vision Research, Rutgers University, Busch Campus-Psychology Building, Piscataway, New Jersey 08854, USA

AN unsolved problem of biology is the processing of global shape in natural vision. The known processes of early vision are spatially restricted (or local) operations, and little is known about their interactions in organizing the visual image into functionally coherent (or global) objects. Here we introduce a human psychophysical method which allows us to measure the effect of perceptual organization on the activity pattern of local visual detectors. We map differential contrast sensitivity for a target across regions enclosed by a boundary. We show that local contrast sensitivity is enhanced within the boundary even for large distances between the boundary and the target. Furthermore, the locations of maximal sensitivity enhancement in the sensitivity maps are determined by global shape properties. Our data support a class of models which describe shapes by the means of a medial axis transformation^{1–3}, implying that the visual system extracts 'skeletons' as an intermediate-level representation of objects. The skeletal representation offers a structurally simplified shape description which can be used for higher-level operations and for coding into memory.

Stimuli were composed of Gabor patches, which are gaussian-modulated sinusoid luminance distributions⁴. Gabor patches model the known receptive-field properties of neurons in the primary visual cortex⁵ (V1). Their oriented shape and small size ensures that the stimuli activate a limited set of early cortical neurons that respond to small parts of the visual field. We measured contrast sensitivity for a Gabor-patch target placed in the context of an enclosing contour and a dense background

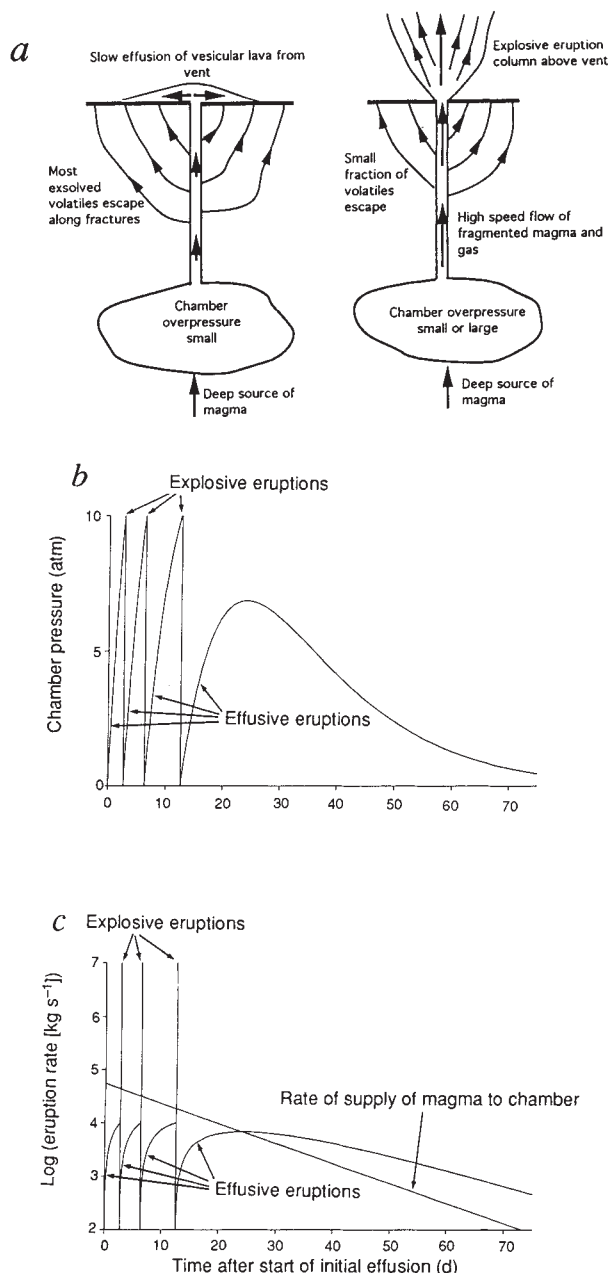
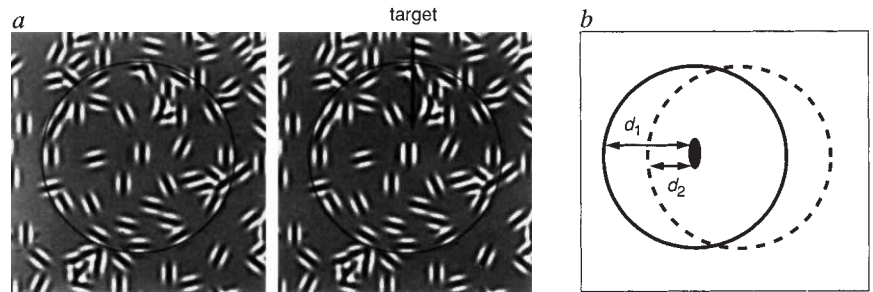


FIG. 4 *a*, Diagram of the effusive and explosive regimes during a model eruption sequence in which chamber pressure (*b*) and eruption rate (*c*) vary according to equation (5) coupled with the conduit flow model (equations (1)–(4); Fig. 3) The deep supply of magma is assumed to decay exponentially¹⁶, $Q_0 = 5.5 \times 10^4 \exp(-10^{-6}t)$ kg s⁻¹, where t is in seconds, the bulk modulus is 10^5 Pa (ref. 19) and the chamber volume is 3×10^{10} m³. The effusive eruption rate (type 'a') is approximated as $10^3 \Delta P$ kg s⁻¹ for chamber overpressures, ΔP , up to a maximum of 10 atm. The explosive eruption rate is 10^7 kg s⁻¹. Three explosive episodes are predicted before the decaying effusive eruption sets in. Further calculations suggest that, owing to the nonlinearity of the process, variation of the deep supply rate can lead to very different detailed predictions of the timing and number of explosive events, although the magnitude and duration of each effusive phase may be similar.

FIG. 1 Stimuli for the contrast sensitivity measurements. *a*, Two sequential frames of one trial (only a small central portion of the stimulus frames is shown). Among the randomly positioned and oriented segments, there is an embedded circle (the continuous circle helps the reader to find the contour). The two frames are equivalent except for the contrast of the central element, which is higher in the second frame. The observer's task was to indicate which frame contained the high-contrast target. *b*, Target–contour distance (d_1 , d_2) was varied during these experiments. The target was presented on the same central location (in the centre of fixation), and the circle was moved closer to or farther from it.

METHODS. Stimuli were composed of band-limited signals (Gabor patches; GP wavelength (λ) was 0.18° ; gaussian envelope size (standard deviation) was equal to λ ; GP amplitude was 16% of mean luminance). The target was a centrally positioned GP with a contrast increment from the background, and its orientation was parallel to the closest segment of an embedded contour. The background contained 2,500 randomly positioned and oriented GPs in a $16^\circ \times 16^\circ$ field. Contrast increment thresholds for the target were measured as a function of target–contour



field (Fig. 1). To avoid uncertainties about the location of the target, it was always presented on the same central location (in the centre of fixation), and the contour was moved around it. Embedded in the random background, the contour segments are not salient locally. This contour becomes salient only when a chain of roughly colinear segments is connected across the gaps^{6,7}.

Determination of the change in contrast sensitivity within a circular contour as a function of contour–target distance revealed a striking enhancement in sensitivity at the centre of the circle (Fig. 2*a*; see also Fig. 6 of ref 7). The enhancement peak in the centre, where contrast sensitivity for the target increased by a factor >2 , was at 1.44° , eight times the Gabor-patch wavelength (λ) distance from the perimeter. The local effect near the contour (peak at $0.36^\circ = 2\lambda$ distance from the perimeter) is consistent with psychophysically measured perceptual spatial interactions^{8,9}. Contrast threshold measurements for a central Gabor target flanked by two Gabor signals showed

distance in a two-alternative temporal forced-choice procedure. One trial consisted of two successive presentations of 170-ms stimulus frames (interframe interval was 500 ms), with the target presented in either the first or second frame. Contrast thresholds were estimated by a staircase procedure¹⁷, four to eight times for each relative target–contour distance. The relative change in sensitivity was calculated by comparing it with the sensitivity for a target with no contour presented. Six human observers participated in the experiments.

that the maximal spatial range of interaction was $5-6\lambda$ in a colinear configuration⁸, but only $2-3\lambda$ when the three Gabor signals were orthogonal to the virtual line connecting them⁹. As the latter case applies to our configuration, the peak at the centre (at 8λ) results from interactions spreading much farther than the known psychophysically measured local interactions. These long-range interactions are specific for the interior of the contour (not shown in Fig. 2*a*; however, in ref. 7, Fig. 6 presents data for the outside region, where sensitivity enhancement peaked at 2λ distance and was zero at 4λ). Analogous results were reported in the primary visual cortex of the macaque monkey, revealing a strong asymmetry in the cells' responses depending on whether the receptive field was positioned inside or outside a figure^{10,11}. These neuronal correlates of figure–ground segregation suggest an early cortical locus for our long-range, interior-specific interactions.

Why does the remote enhancement peak lie at the centre of the circle? To test for the significance of the equidistant property

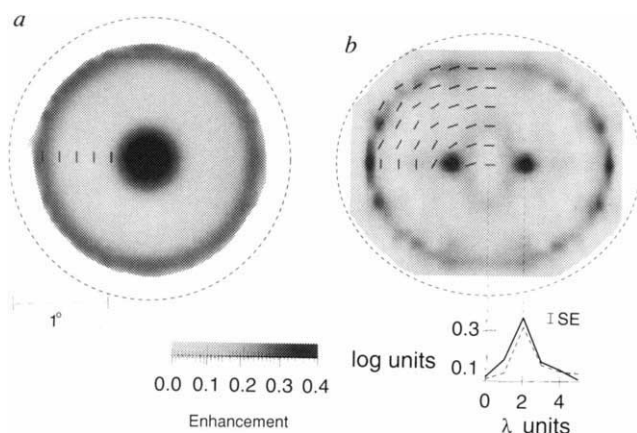
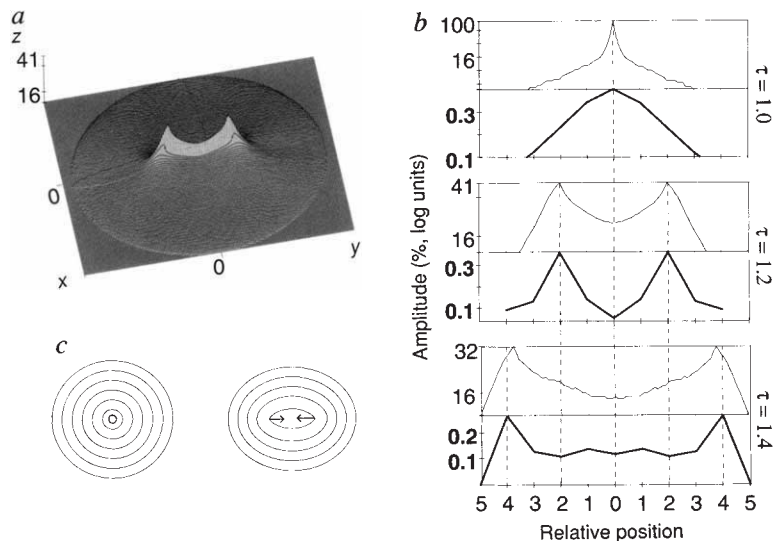


FIG. 2 Maps of contrast sensitivity change for a target within embedded circular and ellipsoid contours. The contours are represented by the dashed lines, and contrast sensitivity change is indicated by different levels of grey, where black stands for sensitivity enhancement >0.3 log units. *a*, Sensitivity change within a circle. The most salient effect was that the circle induced increased sensitivity at its centre, which was at 1.44° distance from the perimeter. Enhancement peak was also observed at 0.36° distance from the contour. The effect of target–contour separation was sampled with 1 target period (λ) resolution (vertical bars represent relative target positions). Averaged results of three observers. (*a* was enhanced from ref. 5, Fig. 6. As we discussed in ref. 5, closed contours are more readily seen than non-closed ones. Here we use perfectly closed contours, and it is not known whether closure is a crucial factor in determining the sensitivity maps.) *b*, Contrast sensitivity change within an ellipse (aspect ratio = 1.2). Note that the centre shows no change in sensitivity, while two peaks can be localized 0.36° away from the centre. Measurements were taken with a resolution of one target period (λ) in one quadrant of the ellipse (superimposed black bars), and the data were reflected symmetrically to the other quadrants to generate the full map. The line graphs show sensitivity change along the major axis for ellipses of different sizes. Ordinate, sensitivity change in log units; abscissa, distance between the target and the centre of the ellipse in λ units. λ was scaled with the size of the ellipse. Dashed line, $\lambda = 0.12^\circ$; 2 observers. Thick line and grey-level map, $\lambda = 0.18^\circ$; 3 observers. Thin line, $\lambda = 0.24^\circ$; 2 observers. Major axes of the ellipses were 16.8λ . Standard error (SE) was averaged across observers. As long as contour–probe distance is expressed in λ units, the overall shape of the sensitivity curves is the same for the different sizes, indicating that the map scales with the figure.

FIG. 3 Peaks of contrast sensitivity can be predicted by the *D*-function. *a*, *D*-function (within an ellipse of 1.2 aspect ratio). *x* and *y* are spatial positions. *D* is defined by the percentage of the boundary points equidistant from (*x*, *y*) within a tolerance of 1% of the boundary length. The two maxima indicate those locations equidistant from the largest portion of the boundary. *b*, Equidistant points predicted by the *D*-function (thin lines) for a circle (aspect ratio, $\tau = 1.0$) and two ellipses ($\tau = 1.2$ and 1.4), and corresponding contrast sensitivity data (thick lines). Abscissa, locations along the major axes (the centre of the conic sections is at 0). Ordinate, percentage of equidistant points on log scale (thin lines) and amplitude of sensitivity change in log units (thick lines). There is very strong conformity between the data and the computed equidistant points. Data are means of 3 observers for each conic section. (The *D*-function was computed by the DAVID visualization software, Fluid Sciences Inc. 1992.) *c*, Propagation of grassfire¹⁻³ within a circle and an ellipse. The grassfire transformation can be imagined as spreading waves of a fire lit along the boundary. Parallels show the succession of fire. A perfect circle collapses into a single point (left panel). A large portion of the waves within an ellipse (right panel) converge in two points along



of the centre, we performed measurements on ellipses. Figure 2*b* shows the sensitivity map of an ellipse (aspect ratio = 1.2). When the centre is not equidistant from all the contour segments there is no enhancement in the centre, and two displaced peaks arise along the major axis (at $0.36^\circ = \lambda$ distance from the centre). The relative peak locations were independent of the absolute size of the ellipse (within a range of 2° – 4° diameter; Fig. 2*b*, line graphs). This indicates that the peak locations do not depend on fixed visual angles or tissue separations in the visual cortex.

Obviously, an ellipse contains no location equidistant from all the boundary points. There are, however, singularities approximately equidistant from a large proportion of the boundary points. Figure 3*a* shows one possible way to estimate these equidistant locations within an ellipsoid contour: at every internal point, a '*D*-function' is defined as the percentage of boundary points equidistant within a tolerance of 1% of the boundary length. The maxima in this space are displaced from the centre along the major axis. Clearly, the *D*-function predicts a single peak within the circle. To test whether the *D*-function can also predict maxima in the ellipse sensitivity maps, we repeated the measurements for an ellipse with 1.4 aspect ratio. Figure 3*b* shows the predicted singularities and the measured maxima for conic sections of 1.0 (circle), 1.2 and 1.4 aspect ratios. Note that the *D*-function predicts sensitivity maxima for the three conic sections within measurement error.

The success of the *D*-function in predicting sensitivity peaks suggests an axis-based shape-processing in human vision. As opposed to conventional boundary tracing, an axis-based shape description is derived in a direction orthogonal to the boundary, where a 'skeleton' (medial axis or 'stick-figure') of a shape is made explicit. Blum's¹⁻³ 'grassfire' transformation was the first biologically motivated implementation of this idea. One way to describe the medial-axis transformation is by analogy to how a

the major axis simultaneously (corners in the central plum-stone), and the residual waves converge later (arrows pointing towards the centre).

grassfire would transverse from the object's boundary into its interior (see Fig. 3*c*). At the last moments, when the 'grass' burns down, the original shape collapses into a skeleton, which consists of quench points equidistant from the boundary. The central peaks in the sensitivity maps we report correspond to the major quench points of the simple, smooth boundaries we used. Small perturbations of more complex boundaries, however, may change the skeleton drastically. To overcome this difficulty, and still provide both the appropriate details and the general form of an object, it has been suggested that the representation should be obtained at various resolutions of different spatial scales¹²⁻¹⁵. Although our sensitivity map data seem to be the first direct evidence for an axis-based shape representation at a given scale, some recent psychophysical data, based on observers' ability to bisect distances between contours, support the multiscale nature of the representation¹⁵.

We have shown for regular two-dimensional planar shapes that local contrast sensitivity is enhanced on particular locations within them. The simplest model that agrees with our results and predicts the locations of maxima in the maps is probably the grassfire model¹⁻³. There are indications that the responsible neural substrates are at the level of the primary visual cortex^{10,11}, but recurrent pathways from higher cortical areas may also participate. As propagation velocities of corresponding cortical areas become known, it will be possible to map our findings from the spatial domain into temporal propagation of neural activity across an extensive medium. Recent findings using real-time optical imaging already provide parameters of a massive cortical activity spread in macaque monkey V1.¹⁶ Our psychophysical findings raise the possibility that long-range lateral interactions behind this activity spread play crucial roles in figure synthesis shape analysis and the extraction of simple spatial relations. □

Received 25 April; accepted 12 July 1994.

- Blum, H. in *Models for the Perception of Speech and Visual Form* (ed. Wathen-Dunn, W.) 362–380 (MIT Press, Cambridge, USA 1967).
- Blum, H. *Perspect. Biol. Med.* **10**, 381–408 (1967).
- Blum, H. *J. theor. Biol.* **38**, 205–287 (1973).
- Gabor, D. *J. IEE Lond.* **93**, 429–457 (1946).
- Marcelja, S. *J. opt. Soc. Am.* **70**, 1297–1300 (1980).
- Field, D. J., Hayes, A. & Hess, R. F. *Vision Res.* **33**, 173–193 (1993).
- Kovács, I. & Julesz, B. *Proc. natn. Acad. Sci. U.S.A.* **90**, 7495–7497 (1993).
- Polat, U. & Sagi, D. *Vision Res.* **33**, 993–999 (1993).
- Polat, U. & Sagi, D. *Vision Res.* **34**, 73–78 (1994).

- Lamme, V. A. F. *Invest. Ophthalm. vis. Sci.* **35**, 1489 (1994).
- Zipser, K., Lee, T. S., Lamme, V. A. F. & Schiller, P. H. *Invest. Ophthalm. vis. Sci.* **35**, 1973 (1994).
- Blum, H. & Nagel, R. N. *Pattern Recogn.* **10**, 167–180 (1978).
- Koenderink, J. J. & van Doorn, A. J. *Biol. Cybern.* **53**, 383–396 (1986).
- Kimia, B. B., Tannenbaum, A. R. & Zucker, S. W. *Int. J. comp. vis.* (in the press).
- Burbeck, C. A. & Pizer, S. M. *Invest. Ophthalm. vis. Sci.* **35**, 1626 (1994).
- Grinvald, A., Lieke, E. E., Frostig, R. D. & Hildesheim, R. J. *Neurosci.* **14**, 2545–2568 (1994).
- Tolhurst, D. & Barfield, L. *Vision Res.* **18**, 951–958 (1978).

ACKNOWLEDGEMENT. We thank Á. Fehér, B. Anderson and J. Nelson for helpful comments.

MHC class II function preserved by low-affinity peptide interactions preceding stable binding

Scheherazade Sadegh-Nasseri*†, Lawrence J. Stern†‡, Don C. Wiley†‡ & Ronald N. Germain*

* Lymphocyte Biology Section, Laboratory of Immunology, NIAID, National Institutes of Health, Bethesda, Maryland 20892, USA

† Department of Biochemistry and Molecular Biology, and the Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA

‡ Present addresses: Department of Immunology, American Red Cross Laboratories, 15601 Crabbs Branch Way, Rockville, MD 20855, USA (S.S.-N.); Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA (L.J.S.).

MAJOR histocompatibility complex class II molecules and their peptide ligands show unusual interaction kinetics, with slow association and dissociation rates that yield an apparent equilibrium constant of $\sim 10^{-6}$ – 10^{-8} M (refs 1–5). However, there is evidence for a specific, rapidly formed, short-lived complex⁶. The altered migration on SDS–polyacrylamide gel electrophoresis of class II molecules upon stable peptide binding^{7–9} has led to the hypothesis that the two kinetically distinguishable types of class II–peptide complexes correspond to different structures. In accord with this model, we demonstrate here that insect cell-derived HLA-DR1

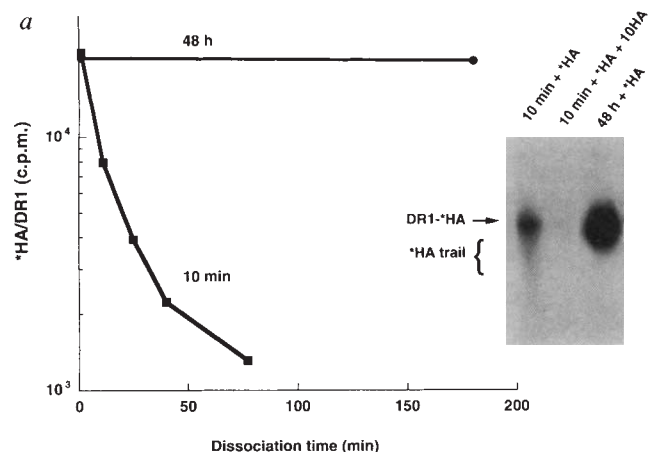
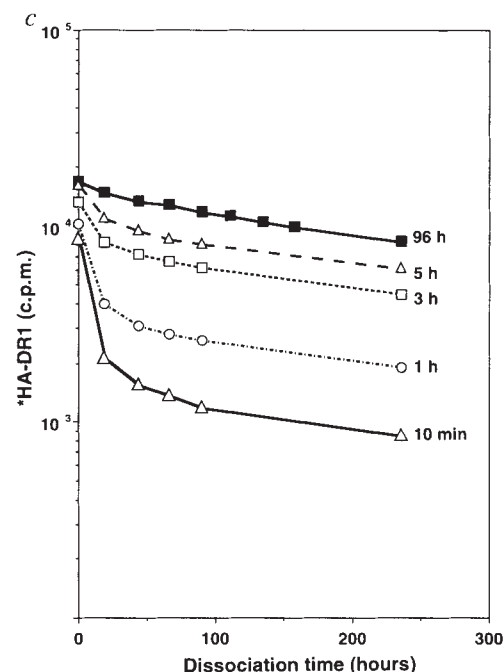
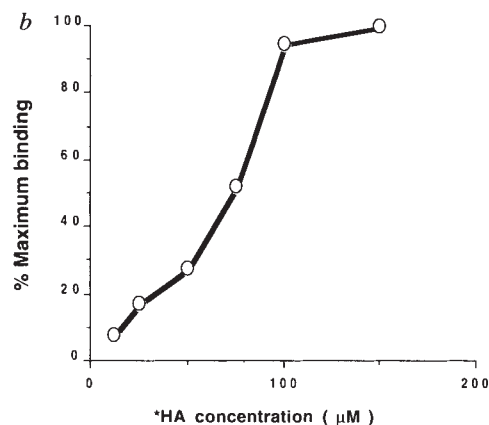


FIG. 1 Empty MHC class II molecules show both fast-on/fast-off and slow-on/slow-off interactions with a single peptide species. **a**, Peptide–DR1 complexes were formed by incubation of 0.3 μ M sDR1 with 100 μ M 125 I-labelled HA peptide (*HA) for 10 min or 48 h at 37 °C. After each incubation, peptide–DR1 complexes were separated from free peptide by spun column filtration and assayed by gamma counting. The kinetics of dissociation of complexes formed during 10 min of incubation were followed by subsequent sequential spun column separations. Inset (**a**), Peptide–DR1 complexes formed by incubation of sDR1 with 100 μ M 125 I-labelled HA peptide for 10 min or 48 h or with 125 I-labelled HA plus a 10-fold excess of unlabelled HA peptide for 10 min at 37 °C were analysed by native gel electrophoresis. **b**, Peptide–DR1 complexes formed by incubation of sDR1 with varying concentrations of 125 I-HA peptide for 10 min at 37 °C and analysed by spun column separation. **c**, Peptide–DR1 complexes were formed by incubation of sDR1 with 125 I-HA peptide for 10 min, 1, 3, 5 and 96 h at 37 °C. The kinetics of dissociation of complexes were measured as in **a**.

METHODS. Soluble DR1 was produced by L243 mAb-affinity purification of supernatant from Sf9 insect cells coinfecting with recombinant baculovirus encoding DR1 α and DR1 β chains truncated just before the transmembrane region⁹ or singly infected with recombinant baculovirus carrying both genes and a dual promoter (L.J.S., unpublished). Purified sDR1 was stored at 1–3 mg ml⁻¹ in phosphate-buffered saline (PBS),



pH 7.2, at 4 °C. Reverse-phase HPLC-purified HA 306–318 peptide (PKYVQNTLKLAT) was labelled with 125 I using Iodobeads (Pierce) and separated from free iodine by passage through a 3-ml G-15 (Pharmacia) gel-filtration column. Fractions containing the labelled peptide were pooled and stored at 4 °C. HPLC analysis showed that the 125 I label coincided with the peptide peak. In most experiments the peptide pool was vacuum-concentrated to 2.5–3 mM ($6\text{--}7 \times 10^6$ c.p.m. per μ l). Peptide concentration was determined by tyrosine absorption or quantitative amino-acid analysis. Reaction mixtures contained sDR1 at 0.3 μ M and, unless otherwise indicated, peptide at 100 μ M in PBS in a total volume of 50 μ l. Spun columns (BioRad; 0.7 ml, containing BioGel P2) were blocked using 20% non-fat dry milk in PBS plus 0.02% Na₂S₂O₃, and used according to the manufacturer's recommendation. There was a linear correlation between the number of fast-forming complexes and the concentration of sDR1 used for incubation (S.S.-N., unpublished observations), which indicates that labelled HA 306–318 binds to sDR1 and not to blocking proteins or other components of the assay system. For measurement of dissociation rates, peptide and sDR1 were mixed and incubated at 37 °C for the indicated times and passed through a spun column at room temperature. The excluded volume was collected, counted, and incubated for the first indicated dissociation time. It was then reappplied to a spun column and the excluded volume recovered and counted. This process was repeated to generate the data for the entire set of dissociation times. Native gel electrophoresis was performed using Mini-PROTEAN II Ready Gels from BioRad as described⁴. Dried gels were autoradiographed.

class II molecules show fast, almost stoichiometric occupancy with rapidly dissociating peptide while remaining sensitive to SDS-induced chain dissociation. The same DR1 molecules slowly and quantitatively form long-lived complexes resistant to SDS-induced denaturation. Surprisingly, low-affinity interaction with peptide protects class II from denaturation at physiological temperature, a finding that has implications for understanding the role of invariant chain in the intracellular behaviour of class II molecules.

Previous experiments in which fast-on/fast-off binding was observed used mouse class II molecules largely occupied with naturally processed peptides⁶. This resulted in measurements reflecting the properties of a small fraction of the class II. Here we examine the binding of a haemagglutinin (HA) peptide (residues 306–318) to soluble DR1 (sDR1) molecules produced by Sf9 insect cells. These secreted class II dimers are free of detectable peptide and bind offered ligand stoichiometrically⁹. Incubation of ¹²⁵I-labelled HA 306–318 with sDR1 for either 10 min or 48 h led to high levels of binding as measured by a rapid-separation technique (Fig. 1a). Fast-forming complexes could also be observed using native gels (Fig. 1a, inset). A relatively high concentration of peptide (100 μ M) was required for rapid saturation binding (Fig. 1b), and the resulting complexes had a $t_{1/2\text{off}} \sim 10$ min (Fig. 1a), consistent with a low-affinity, fast-on/fast-off interaction. In contrast, sDR1–HA complexes generated at longer incubation times showed an increasing proportion of long-lived forms, whose $t_{1/2\text{off}}$ was ~ 140 h at 37 °C (Fig. 1c). The complexes generated upon prolonged incubation showed

occupancy approaching 100%. Owing to their rapid dissociation rate, small changes in separation time led to greater variability in the measured number of rapidly forming complexes, ranging over 50–90% of maximum binding. These data are consistent with the same DR1 molecule binding peptide in either of two states.

Rapid peptide binding is specific. sDR1 showed rapid binding to HA 306–318 but not to a moth cytochrome *c* peptide (residues 88–103), whereas this latter peptide but not HA showed fast-on/fast-off binding to purified $E\alpha^kE\beta^k$ (Fig. 2a). This also demonstrates that fast-on/fast-off binding is not unique to the sDR1–HA combination, and occurs with class II molecules produced by normal B cells coexpressing invariant chain. Cold HA 306–318 competes for labelled peptide binding in direct proportion to its concentration (Fig. 2b) and incubation of sDR1 with mixtures of labelled and unlabelled HA having a constant total peptide concentration give a linear binding relationship with a slope of 1 with respect to the concentration of labelled peptide for both types of complexes (Fig. 2c). These latter data argue strongly against the possibility that peptide–DR1 complexes with different off rates arise from heterogeneous iodination of HA 306–318 at the tyrosine involved in high-affinity binding^{4,10,11}.

Stable binding of peptide to class II heterodimers induces a change in the biochemical properties of the major histocompatibility complex (MHC) molecules^{8,9,12,13}. Increased resistance to subunit dissociation during SDS–PAGE is a convenient assay

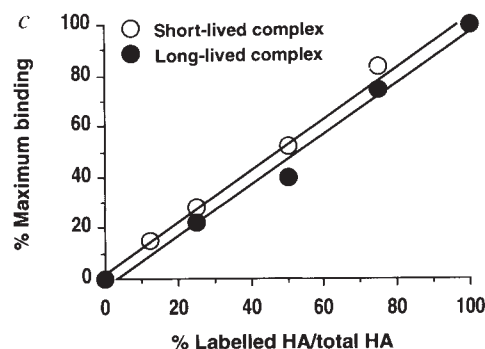
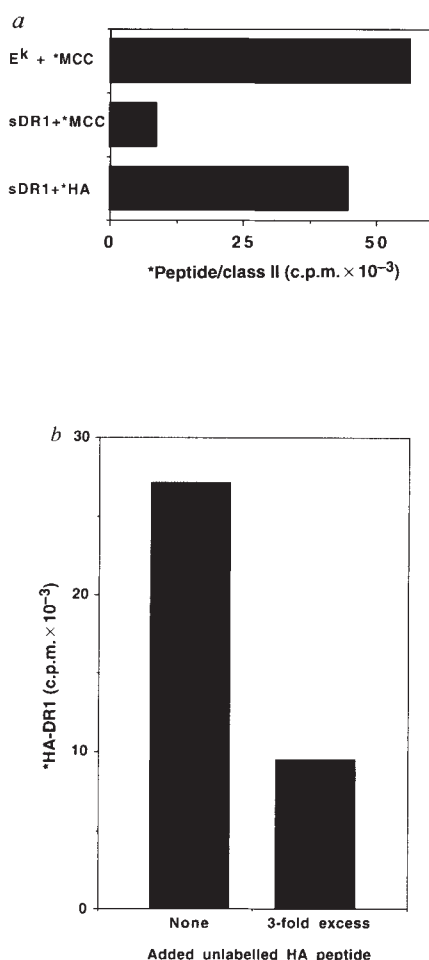
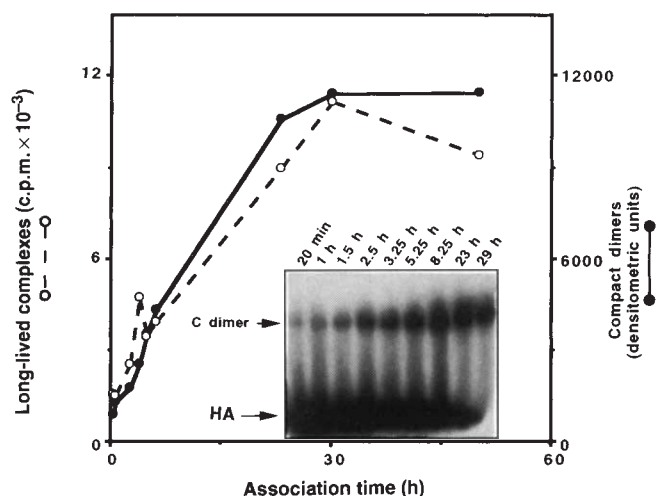


FIG. 2 Both fast- and slow-forming sDR1–peptide complexes are specific and involve the same derivatized peptide species. **a**, Fast-forming complexes show peptide and MHC molecule specificity. sDR1 or $E\alpha^kE\beta^k$ preparations enriched for molecules with empty binding sites were incubated with ¹²⁵I-HA or ¹²⁵I-labelled moth cytochrome *c* for 10 min, followed by spun-column separation. Data represent c.p.m. in the excluded volume. **b**, Unlabelled HA peptide competes proportionately with ¹²⁵I-HA for rapid binding to sDR1. **c**, The different kinetic forms of sDR1/HA–peptide complexes do not arise from interactions involving two distinct peptide species with different binding affinities.

METHODS. $E\alpha^kE\beta^k$ was prepared from detergent lysates of CBA/J spleen cells by affinity purification using a 14-4-4S mAb column as described. $E\alpha^kE\beta^k$ with a high proportion of empty sites was isolated as the first fraction eluting from the affinity column²². These molecules migrate as floppy or unstable dimers on SDS–PAGE. Moth cytochrome *c* peptide 88–103 (ANERADLIAYLKQATK) was iodinated as described for HA peptide in Fig. 1 legend. Binding reactions and spun column separations are described in Fig. 1 legend, except that for **b**, the indicated concentration of unlabelled HA peptide was added simultaneously with 150 μ M ¹²⁵I-HA and binding measured after 10 min; and for **c**, varying volumes of ¹²⁵I-HA (0.5, 1, 2 and 4 μ l) were mixed with 3.5, 3, 2 and 0 μ l unlabelled HA each at 1 mM, then added to 36 μ l sDR1 solution to yield 100 μ M (final) total peptide in each case. After incubation for 10 min or 48 h at 37 °C, incubation mixtures were subjected to spun-column separation, and the excluded volume was counted. The figure displays the relationship between the concentration of labelled peptide and extent of complex formation.



for this change, although this pattern of electrophoretic behaviour is not seen with all long-lived class II-peptide combinations^{12,14,15}. Brief incubation with unlabelled HA did not induce any detectable change in the migration pattern of sDR1 in SDS-PAGE, as both empty and rapidly loaded DR1 migrated as dissociated chains under these conditions (data not shown). A kinetic analysis of stable (compact) dimer and long-lived complex formation is shown in Fig. 3. The total labelled peptide signal from HA-sDR1 complexes (short lived plus long lived) at different incubation times was relatively constant. Formation of compact dimers (Fig. 3, solid circles and inset) and long-lived peptide-sDR1 complexes (Fig. 3, open circles) occurred at the same rate, with a $t_{1/2}$ on ~ 7 h at 37°C . As more compact dimers formed, fewer short-lived complexes were detected, as indicated by a smaller free peptide signal at the dye front (Fig. 3, inset). Thus, most or all sDR1 molecules have binding sites able to form complexes with peptide rapidly, but generation of long-lived complexes occurs in parallel with an infrequent structural change that takes place in the population of peptide-interacting sDR1 molecules with a $t_{1/2}$ of 6–7 h.

While investigating stable complex formation at varying peptide concentrations, we noted that each concentration led to different plateau binding, despite the use of at least a 20-fold molar excess of peptide over sDR1 (Fig. 4a). Furthermore, once long-lived complex formation approached a plateau at a sub-saturating peptide concentration, addition of a saturating concentration of ligand did not yield the predicted number of additional long-lived complexes (Fig. 4b), suggesting that most sDR1 molecules that had not formed stable complexes by this

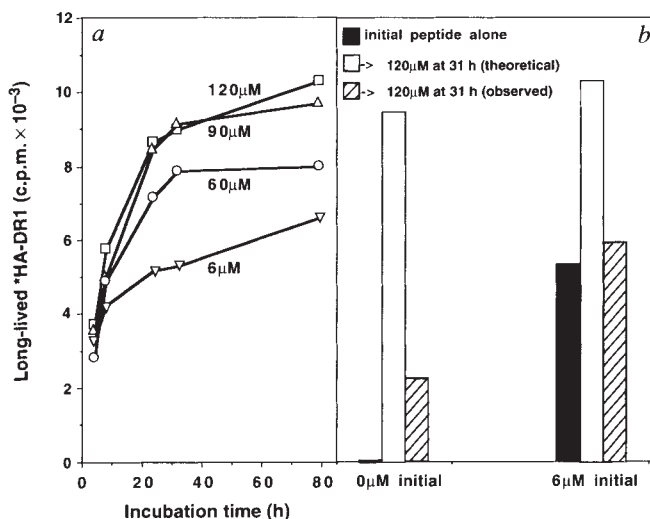
time had become functionally inactive or denatured. The plateaux of stable complex formation roughly correlated with the levels of fast-dissociating complexes generated at each peptide concentration (Fig. 1b), implying that formation of rapidly dissociating complexes preserves class II from binding site inactivation at physiological temperature.

These experiments provide strong evidence for two distinct kinetic modes of interaction involving a single peptide-MHC class II pair. The demonstration of both binding states using distinct peptides and different class II molecules from humans and mice suggests that these two binding states may be a common, though not always readily detectable, characteristic of class II-peptide reactions. Our results are also important in that they demonstrate the near-stoichiometric occupancy of the same population of class II molecules with either rapidly or slowly dissociating peptide. This rules out the possibility that fast-on/fast-off binding reflects the behaviour of a subset of class II molecules structurally incapable of stable peptide acquisition. Experiments with living cells suggest that fast-on/fast-off binding can be physiologically relevant¹⁶. Our data also make clear that the slow formation of long-lived class II-peptide complexes can be only partially accounted for by the gradual generation of new binding sites from class II molecules preoccupied with naturally processed peptides¹⁷, as sDR1 molecules whose binding sites are uniformly accessible to peptide in as little as 10 min still show a forward $t_{1/2}$ of 5–7 h for formation of stable complexes at neutral pH.

Formation of long-lived complexes occurs at the same rate as a change in class II biochemical behaviour, revealed for the

FIG. 4 Peptide concentration-dependent survival of sDR1 molecules at 37°C . a, sDR1 was incubated at 37°C with ^{125}I -HA peptide (0, 6, 60, 90 and $120\ \mu\text{M}$) for the times indicated. Long-lived peptide-sDR1 complexes were quantified for each incubation time by two sequential spun-column separations (Fig. 3 legend). b, Samples incubated for 31 h with 6 or $0\ \mu\text{M}$ peptide at 37°C were each split into two tubes and additional peptide was added to one sample of each group to raise the final concentration of peptide to $120\ \mu\text{M}$. These samples were further incubated at 37°C for 48 h and then analysed by sequential spun-column separation for long-lived complex formation.

METHODS. A cocktail of protease inhibitors was included in the sDR1 preparation⁹ to inhibit proteolysis, and no significant class II α - or β -chain degradation was observed under these incubation conditions as assayed by SDS-PAGE or native gel electrophoresis.



sDR1-HA combination as enhanced stability during SDS-PAGE. Thus, in accord with the kinetic-intermediate model⁶, the slow event in peptide binding correlates with a change in class II dimer structure. The rare nature of this structural change among rapidly occupied class II molecules might be due either to an intrinsic energetic barrier for conversion of class II to this higher-affinity state, or to the low frequency with which an interacting peptide is optimally oriented in the binding groove. The latter possibility is consistent with the chemistry of HA peptide interaction with DR1 molecules that involves conserved hydrogen bonds between class II and the main-chain atoms of the peptide, as well as many peptide side-chain/DR1 interactions¹⁰. These properties could permit class II to interact weakly with a peptide whose side-chain anchors do not all sit properly in pockets within the binding groove, and suggest that stable binding occurs when the side chains and pockets are aligned properly.

An intriguing finding is the ability of low-affinity peptide occupancy to preserve class II from the inactivation at physiological temperature. We and others have reported that invariant chain is important in stabilizing the association of class II α - and β -subunits within living cells without yielding the SDS-resistant phenotype^{18–20}. The ability of low-affinity peptide binding to preserve class II heterodimers without promoting their resistance to SDS denaturation raises the possibility that a similar quality of binding site occupancy by a segment of intact invariant chain helps maintain the function of class II

molecules until they arrive at an endosomal site appropriate for antigen capture^{21,22}. □

Received 5 April; accepted 15 July 1994.

1. Babbitt, B. P., Allen, P. M., Matsueda, G., Haber, E. & Unanue, E. R. *Nature* **317**, 359–361 (1985).
2. Buus, S., Sette, A., Colon, S. M., Jenis, D. M. & Grey, H. M. *Cell* **47**, 1071–1077 (1986).
3. Roche, P. A. & Cresswell, P. *J. Immunol.* **144**, 1849–1856 (1990).
4. Jardetzky, T. S. et al. *EMBO J.* **9**, 1797–1803 (1990).
5. Roof, R. W., Leuschner, I. F. & Unanue, E. R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1735–1739 (1990).
6. Sadegh-Nasseri, S. & McConnell, H. M. *Nature* **337**, 274–276 (1989).
7. Wettstein, D. A., Boniface, J. J., Reay, P. A., Schild, H. & Davis, M. M. *J. exp. Med.* **174**, 219–228 (1991).
8. Sadegh-Nasseri, S. & Germain, R. N. *Nature* **353**, 167–170 (1991).
9. Stern, L. J. & Wiley, D. C. *Cell* **68**, 465–477 (1992).
10. Stern, L. J., Brown, J. H., Jardetzky, T. S. & Wiley, D. C. *Nature* **368**, 215–221 (1994).
11. Hammer, J. et al. *Cell* **74**, 197–203 (1993).
12. Germain, R. N. & Rinker, A. G. Jr *Nature* **363**, 725–728 (1993).
13. Kozono, H., White, J., Clements, J., Marrack, P. & Kappler, J. *Nature* **369**, 151–154 (1994).
14. Lanzavecchia, A., Reid, P. A. & Watts, C. *Nature* **357**, 249–252 (1992).
15. Nelson, C. A., Petzold, S. J. & Unanue, E. R. *Proc. natn. Acad. Sci. U.S.A.* **90**, 1227–1231 (1993).
16. Fairchild, P. J., Wildgoose, R., Atherton, E., Webb, S. & Wraith, D. C. *Int. Immun.* **5**, 1151–1158 (1993).
17. Tampe, R. & McConnell, H. M. *Proc. natn. Acad. Sci. U.S.A.* **88**, 4661–4665 (1991).
18. Viville, S. et al. *Cell* **72**, 635–648 (1993).
19. Bikoff, E. K. et al. *J. exp. Med.* **177**, 1699–1712 (1993).
20. Bonnerot, C. et al. *EMBO J.* **13**, 934–944 (1994).
21. Romagnoli, P. & Germain, R. N. *J. exp. Med.* (in the press).
22. Sadegh-Nasseri, S. in *Antigen Processing and Presentation* (eds Humphreys, R. E. & Pierce, S. K.) 170–187 (Academic, San Diego, 1994).

ACKNOWLEDGEMENTS. S.S.N. was supported in part by Oravax, Inc. L.J.S. acknowledges a fellowship from the Daymon Runyon-Walter Winchell Cancer Institute; D.C.W. is an investigator with the Howard Hughes Medical Institute. We thank D. Margulies for discussion and for reviewing the manuscript.

Cytolytic T-cell cytotoxicity is mediated through perforin and Fas lytic pathways

Bente Lowin, Michael Hahne, Chantal Mattmann & Jürg Tschopp*

Institute of Biochemistry, University of Lausanne, CH-1066 Epalinges, Switzerland

THE recent generation of perforin knock-out mice^{1,2} has demonstrated a crucial role for the pore-forming perforin in cytolytic T-lymphocyte (CTL)-mediated cytotoxicity. Perforin-deficient mice failed to clear lymphocytic choriomeningitis virus *in vivo*, yet substantial killing activity still remained in perforin-free CTLs *in vitro*, indicating the presence of (a) further lytic pathway(s). Fas is an apoptosis-signalling receptor molecule on the surface of a number of different cells. Here we report that both perforin-deficient and Fas-ligand-deficient CTLs show impaired lytic activity on all target cells tested. The killing activity was completely abolished when both pathways were inactivated by using target cells from Fas-receptor-deficient *lpr* mice and perforin-free CTL effector cells. Fas-ligand-based killing activity was triggered upon T-cell receptor occupancy and was directed to the cognate target cell. Thus, two complementary, specific cytotoxic mechanisms are functional in CTLs, one based on the secretion of lytic proteins and one which depends on cell-surface ligand-receptor interaction.

The analysis of perforin-deficient mice has revealed the important role of perforin in CTL-mediated cytotoxicity^{1,2}. While Kägi *et al.* concluded from their results that perforin accounts for 100% of the lytic activity against most target cells¹, we measured substantial lysis of all haematopoietic target cells, when apoptosis rather than plasma membrane damage was taken as an indicator for cell death². Moreover, using adherent instead of

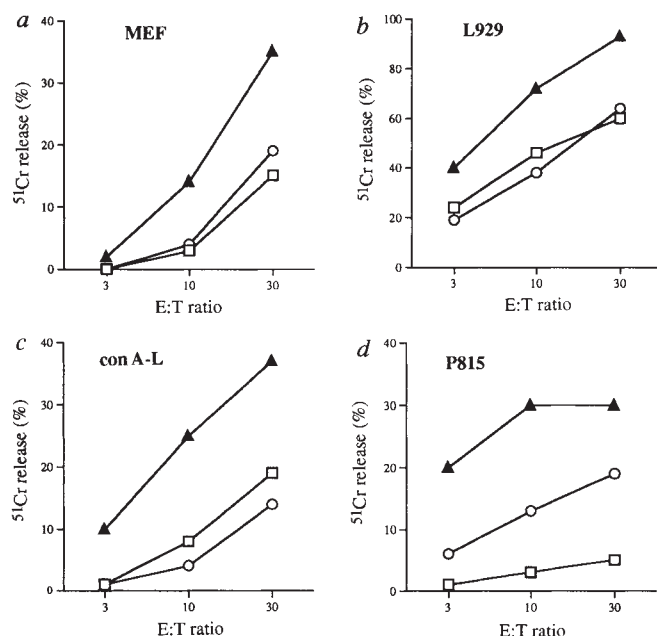
trypsin-detached cells, one-third of the killing activity remained² with fibroblasts as target cells. To corroborate the presence of alternative perforin-independent killing pathways, major histocompatibility complex (MHC) alloantigen-specific cytotoxicity of C57BL/6 (H-2^b) anti-MRL (H-2^k) cultures was assayed on additional target cells, including mouse primary embryonic fibroblasts, the L929 fibroblast cell line and concanavalin A (conA)-stimulated lymphocytes. Although cytotoxicity was impaired in the absence of perforin, all target cells were considerably lysed (Fig. 1a–c). The residual lytic activity persisted in perforin-deficient CTLs, when lysis was redirected to nonspecific targets via T-cell-receptor-specific antibodies (Fig. 1d).

The cell-surface receptor Fas can transduce signals which lead to apoptotic cell death^{3,4}. The Fas-based death pathway plays an important regulatory role *in vivo*. *lpr* mice, devoid of functional Fas receptor (FasR)⁵, as well as *gld* mice, lacking the active Fas ligand (FasL)^{6,7}, suffer from similar auto-immune pathologies, including the accumulation of large numbers of lymphocytes in spleen and lymph nodes. Fas also contributes to nonspecific CTL cytotoxicity observed after polyclonal activation of lymphocytes^{8,9}.

To evaluate the contribution of Fas in specific CTL cytotoxicity, mixed lymphocyte cultures of *gld*-derived spleen cells (FasL[−]) were generated. Fibroblasts as well as conA-stimulated lymphoblasts were lysed by FasL-deficient lymphocytes, albeit with reduced activity compared to wild-type CTLs (Fig. 1a–d). The cytolytic activity was comparable to that detected with perforin-deficient CTLs. About 10–30% of killing activity was left in each case. Experiments were now devised whereby both pathways were functionally inactive (Fig. 2). Target cells were derived from the C3H/*lpr* mice (FasR[−]). Normal CTLs (perforin⁺/FasL⁺) as well as *gld*-derived CTLs (FasL[−]) still showed lytic activity (presumably perforin-based). In contrast, perforin-deficient CTLs were completely inactive against FasR-lacking primary fibroblasts or lymphocyte targets. Thus, Fas- and perforin-dependent pathways are responsible for all detectable killing activity in these experimental systems.

Target-cell recognition and killing by CTLs is highly specific, sparing bystander cells, that is, non-cognate targets mixed

* To whom correspondence should be addressed.



among cognate targets¹⁰. Exocytosis of granule components occurs highly vectorially towards the killer-target cell surface^{11,12}, accounting for the specificity of the perforin pathway. In contrast, nonspecific activation of lymphocytes^{7,8,13} or overexpression of the Fas ligand *in vitro*¹⁴ can lead to nonspecific Fas-dependent destruction of target cells. Unrestrained killing of all FasR-bearing cells by CTLs would, however, have deleterious consequences *in vivo* as a result of widespread cellular distribution of the FasR¹⁵. We therefore assessed the specificity of the FasL-based lytic mechanism of CTLs. The killing activity of perforin-deficient (intact Fas pathway) allogeneic CTLs (H-2^b anti-H-2^d) was tested against cognate H-2^d (3T3 fibroblast) and irrelevant H-2^k (H-2^k embryonic fibroblast, H-2^b syngeneic stimulated splenocytes) target cells (Fig. 3). After 4–8 h incubation, high cytolytic activity was detected only on cognate H-2^d target cells, even in the presence of phorbol 12-myristate 13-acetate (PMA)/ionomycin, known to upregulate FasL expression^{8,14}, or T-cell-receptor-stimulating cognate target cells (Fig. 3, and data not shown). Only in a later phase (16 h) did

FIG. 1 Killing of various target cells by perforin- and FasL-deficient CTLs. MHC alloantigen-specific T cells (H-2^b anti H-2^k) were generated in 5-day mixed lymphocyte cultures *in vitro* from normal control mice (filled triangles), perforin (–/–) mice (open squares) and FasL-deficient *gld* mice (open circles). The CTL activity was tested on MEF (C3H mouse-derived untransformed embryonic fibroblasts, H-2^k) (a), L929 (fibrosarcoma, H-2^k) (b) and conA-L (conA-activated spleen cells, H-2^k) (c). d, The cytotoxic activity was redirected to P815 (mastocytoma, H-2^d) with anti-T-cell-receptor antibodies bound to Fc receptors of the P815 cells (H-2^b). The lytic activities were determined in ^{51}Cr -release assays. METHODS. 129 Perforin knock-out mice (H-2^b) were generated as described previously² and crossed into C57BL/6 mice (H-2^b). C57BL/6 *gld* mice were obtained from The Jackson Laboratory (Bar Harbor, ME, U.S.A.). Mouse embryonic fibroblasts were prepared from 12–14-day-old embryos. For conA blasts, spleen cells were freed from red blood cells and activated with conA ($4\text{ }\mu\text{g ml}^{-1}$) for 3 days^{8,17}. All primary target cells were isolated from MRL mice (H-2^k). Alloreactive cytotoxic T lymphocytes were generated in a 5-day mixed lymphocyte culture¹⁸. Responder spleen cells ($2 \times 10^6\text{ ml}^{-1}$) were from adult (6–8 week-old) mice with the normal perforin gene (+/+), mice homozygous (–/–) for the disrupted perforin gene and *gld* mice. All these mice were of haplotype H-2^b. Stimulators were irradiated spleen cells ($5 \times 10^6\text{ ml}^{-1}$, 3,000 rad) from CBA (H-2^k) mice. Cellular cytotoxicity was assayed after 4 h for haematopoietic targets and after 8 h for fibroblasts¹⁹.

Fas-dependent T-cell-mediated lysis become less focused, permitting nonspecific destruction of non-adhering neighbouring cells (B.L. *et al.*, unpublished results). The high initial specificity of the CTL–Fas pathway raises the question of the mechanisms underlying the control of this activity.

In all experiments shown here, both the perforin- and Fas-dependent pathways were always implicated in T-cell-mediated cytotoxicity and accounted for 100% of detectable killing activity. Based on the lysis of the limited selection of target cells shown in Fig. 2, we estimate that the perforin mechanism is responsible for approximately two-thirds of the killing activity, one-third being contributed by Fas *in vitro*. This ratio probably varies *in vivo*. The FasR expression on cells differs³ or may

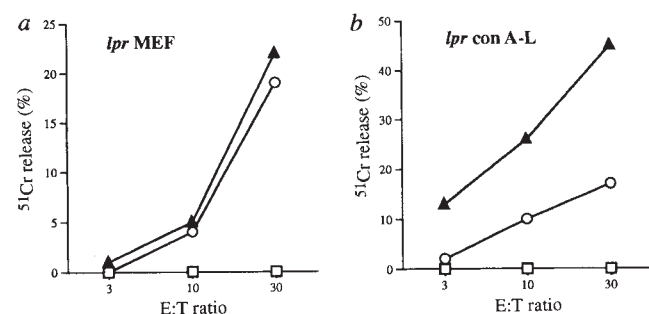


FIG. 2 The cytotoxicity of mixed lymphocyte cultures on FasR-negative target cells. The killing activity of MHC alloantigen-specific CTLs (H-2^b anti-H-2^d, see Fig. 1) derived from control mice (closed triangles), from perforin (–/–) mice (open squares) and FasL-deficient *gld* mice (open circles) was tested on lpr MEF (MRL/lpr mouse-derived untransformed embryonic fibroblasts, H-2^k) (a), and lpr conA-L (conA-activated spleen cells from MRL/lpr mice, H-2^k) (b). METHODS. 4–6 week-old MRL and MRL/lpr mice were obtained from H. Olac. Target cell preparation and cytotoxicity assays are described in Fig. 1 legend.

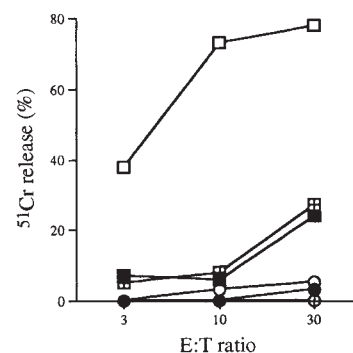


FIG. 3 Specificity of CTL Fas-mediated cytotoxicity. Secondary mixed lymphocyte cultures (H-2^b against H-2^d) were generated from perforin-deficient mice and tested for their lytic activity against specific H-2^d-bearing 3T3 fibroblasts (squares) or nonspecific H-2^k-bearing MEFs (circles). Cytotoxic activity was assayed in medium alone (filled symbols) or, in order to upregulate the expression of the FasL, in the presence of either an equal number of cold specific target cells (P815, H-2^d) (crossed symbols) or PMA and ionomycin (open symbols). Similar results were obtained when P815 mastocytoma and third-party H-2^b syngeneic stimulated splenocytes were taken (data not shown). Cytotoxicity was measured after 4 h.

METHODS. Cytotoxicity tests were performed as described for Fig. 1. Either 10^4 cold, cognate P815 target cells per well containing 10^4 labelled target cells or a mixture of PMA (final concentration 5 ng ml^{-1}) and ionomycin (final concentration $0.5\text{ }\mu\text{g ml}^{-1}$) were added during the 4-h incubation period in some experiments.

be modulated by pathogens; alternatively, perforin and FasL content may vary in CTL effector cells. Downregulation of the FasR on cells infected by viruses may explain why perforin-deficient mice fail to clear lymphocytic choriomeningitis virus¹, in spite of the presence of a functional Fas pathway. Alternatively, FasL/perforin double-deficient mice may have greater difficulty in eliminating the virus than the perforin knock-out mice. This would indicate that the Fas pathway is functional, but is not sufficient to clear this virus in the absence of perforin.

The two pathways are complementary, one requiring cell-cell contact and using a regulated ligand able to lyse only Fas-

receptor-bearing cells, the second a perforin-granzyme-based mechanism whose lytic activity is dependent on the presence of phosphorylcholine¹⁶, allowing virtually any target cell to be killed. Although both pathways induce apoptosis, this intracellular killing mechanism prevails in the Fas-mediated pathway, whereas relative membrane damage is higher in the perforin killing mechanism (B.L. *et al.*, unpublished results). These two cytolytic systems may have converged during evolution in the same cytolytic cell to equip the cytolytic cell optimally against the multitude of escape strategies used by invading pathogens. □

Received 21 June; accepted 1 August 1994.

- Kägi, D. *et al.* *Nature* **369**, 31–37 (1994).
- Lowin, B., Beermann, F., Schmidt, A. & Tschopp, J. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
- Itoh, N. *et al.* *Cell* **66**, 233–243 (1991).
- Trauth, B. C. *et al.* *Science* **245**, 301–305 (1989).
- Itoh, N., Tsujimoto, Y. & Nagata, S. *J. exp. Med.* **178**, 461–468 (1993).
- Takahashi, T. *et al.* *Cell* **76**, 969–976 (1994).
- Ramsdell, F. *et al.* *Eur. J. Immun.* **24**, 928–933 (1994).
- Vignaux, F. & Golstein, P. *Eur. J. Immun.* **24**, 923–927 (1994).
- Rouvier, E., Luciani, M. F. & Golstein, P. *J. exp. Med.* **177**, 195–200 (1993).
- Cerottini, J. C., Engers, H. D., MacDonald, H. R. & Brunner, T. J. *J. exp. Med.* **140**, 703–717 (1974).
- Kupfer, A., Singer, S. J. & Dennert, G. *J. exp. Med.* **163**, 489–498 (1986).

- Peters, P. *et al.* *Eur. J. Immun.* **19**, 1469–1475 (1989).
- Russell, J. H., Rush, B., Weaver, C. & Wang, R. *Proc. natn. Acad. Sci. U.S.A.* **90**, 4409–4413 (1993).
- Suda, T., Takahashi, T., Golstein, P. & Nagata, S. *Cell* **75**, 1169–1178 (1993).
- Ogasawara, J. *et al.* *Nature* **364**, 806–809 (1993).
- Tschopp, J., Schafer, S., Masson, D., Peitsch, M. C. & Heusser, C. *Nature* **337**, 272–274 (1989).
- Colligan, J. E. *Current Protocols in Immunology* 3121–3126 (Wiley-Interscience, New York, 1992).
- Maryanski, J. L., MacDonald, H. R. & Cerottini, J. C. *J. Immun.* **124**, 42–48 (1980).
- Ucker, D. S. *et al.* *Molec. cell. Biol.* **12**, 3060–3069 (1992).

ACKNOWLEDGEMENTS. We thank C. Kamel, R. MacDonald, F. Beerman and J. Maryanski for helpful discussions and comments. This work was supported by the Swiss National Science Foundation.

Inhibition of endocytosis by elevated internal calcium in a synaptic terminal

Henrique von Gersdorff & Gary Matthews

Department of Neurobiology and Behavior,
State University of New York, Stony Brook,
New York 11794-5230, USA

DURING synaptic transmission in the nervous system, synaptic vesicles fuse with the plasma membrane of presynaptic terminals, releasing neurotransmitter by exocytosis^{1,2}. The vesicle membrane is then retrieved by endocytosis and recycled into new transmitter-containing vesicles^{3–6}. Exocytosis in synaptic terminals is calcium-dependent^{7–9}, and we now report that endocytosis also is regulated by the intracellular calcium concentration ($[Ca^{2+}]_i$). Capacitance measurements^{10,11} in synaptic terminals of retinal bipolar neurons revealed that endocytosis was strongly inhibited by elevated $[Ca^{2+}]_i$ in the range achieved by Ca^{2+} -current activation. The rate of membrane retrieval was steeply dependent on $[Ca^{2+}]_i$, with a Hill coefficient of 4 and half-inhibition at ~ 500 nM. At $[Ca^{2+}]_i \geq 900$ nM, endocytosis was entirely absent. The action of internal calcium on endocytosis represents a novel negative-feedback mechanism controlling the rate of membrane recovery in synaptic terminals after neurotransmitter secretion. As membrane retrieval is the first step in vesicle recycling, this mechanism may contribute to activity-dependent synaptic depression.

Figure 1a shows the capacitance change in a bipolar-cell synaptic terminal following brief (250 ms) activation of calcium current, which elevated $[Ca^{2+}]_i$ transiently to ~ 400 nM. Capacitance increased rapidly (exocytosis) and then recovered exponentially to baseline (endocytosis) without delay after the stimulus. Recovery following 10-ms pulses, which elicited ~ 10 -fold smaller capacitance jumps, occurred at a similar rate (average exponential time constant, $\tau = 1.4$ s; 7 cells), indicating that the rate of retrieval did not depend on the degree of preceding exocytosis. In contrast, after longer depolarization (1 s) had elevated $[Ca^{2+}]_i$ to > 1 μ M, the capacitance remained at an elevated plateau until $[Ca^{2+}]_i$ fell to ~ 700 nM (Fig. 1b), suggesting that elevated

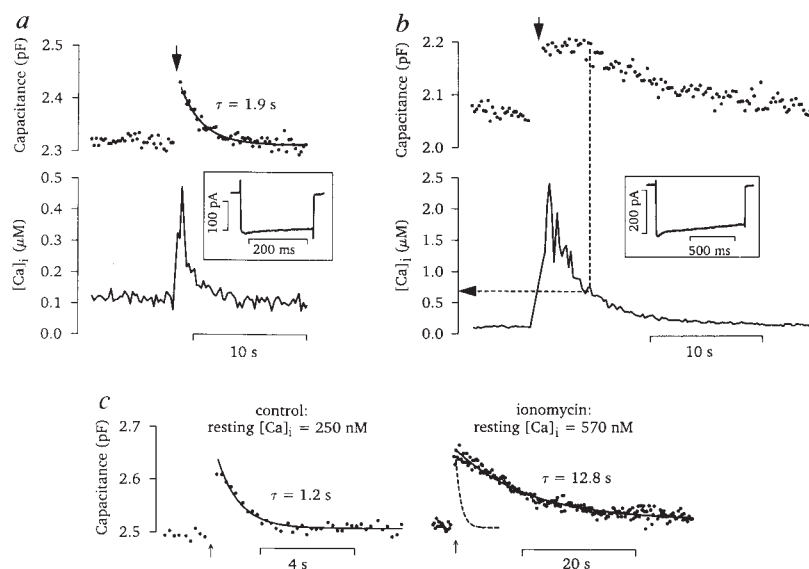
$[Ca^{2+}]_i$ inhibits endocytosis. To test this, endocytosis was compared before and after $[Ca^{2+}]_i$ had been increased with ionomycin (Fig. 1c). At resting $[Ca^{2+}]_i$, retrieval was rapid, with $\tau = 1.2$ s. When $[Ca^{2+}]_i$ was increased with ionomycin, τ slowed to 12.8 s. Thus, elevated $[Ca^{2+}]_i$ (in a range achievable by activation of Ca^{2+} current, that is, ~ 1 μ M) slows the rate of endocytosis following exocytosis ($[Ca^{2+}]_i$ in this range does not stimulate exocytosis⁹). This contrasts with previous results in neuroendocrine cells. Endocytosis in pituitary melanotrophs is little affected by $[Ca^{2+}]_i$ in the micromolar range¹², whereas higher $[Ca^{2+}]_i$ (50–500 μ M) stimulates endocytosis in melanotrophs¹³, adrenal chromaffin cells¹¹ and pituitary nerve terminals¹⁴. Synaptic terminals and neuroendocrine cells thus differ markedly in Ca^{2+} regulation of endocytosis.

To determine the range of $[Ca^{2+}]_i$ at which endocytosis is inhibited, we dialysed synaptic terminals with solutions containing Ca^{2+} buffered to various levels (Fig. 2a–c). Endocytosis occurred at $[Ca^{2+}]_i$ up to 800–900 nM, and the retrieval rate was graded with $[Ca^{2+}]_i$. At the highest levels, endocytosis ceased altogether (Fig. 2c). Figure 2d summarizes the relationship between $[Ca^{2+}]_i$ and the rate constant of endocytosis, α , estimated by fitting an exponential to the recovery phase of the capacitance response. The smooth curve was drawn according to the Hill equation: $\alpha = \alpha_{\max}(1 - 1/(1 + (K_{1/2}/[Ca^{2+}]_i)^n))$, where $\alpha_{\max}$ (0.54 s⁻¹) is the asymptotic rate of endocytosis at low $[Ca^{2+}]_i$, $K_{1/2}$ (460 nM) is $[Ca^{2+}]_i$ producing half-inhibition, and n (4.0) is the Hill coefficient. Figure 2e shows that, unlike the endocytotic rate, the average amplitude of exocytosis did not consistently change with $[Ca^{2+}]_i$ in this range.

We used the Hill equation of Fig. 2d and the measured time course of $[Ca^{2+}]_i$ after stimulation to predict the time course of individual capacitance responses like that in Fig. 1b, which showed delayed recovery of capacitance after stimulation. An exponential decay was fitted to the calcium response, yielding $[Ca^{2+}]_i$ as a function of time, which was then used in the Hill equation (above) to calculate the rate constant for endocytosis as a function of time. Figure 3 compares actual (filled circles) and calculated (smooth curve) capacitance responses. The fit seems adequate given that there are no free parameters after fitting the $[Ca^{2+}]_i$ response, with values for $\alpha_{\max}$, $K_{1/2}$ and n being taken from Fig. 2d. The mean square error (see Fig. 3 legend) of the fit was 1.8×10^{-5} in Fig. 3; similar fits were

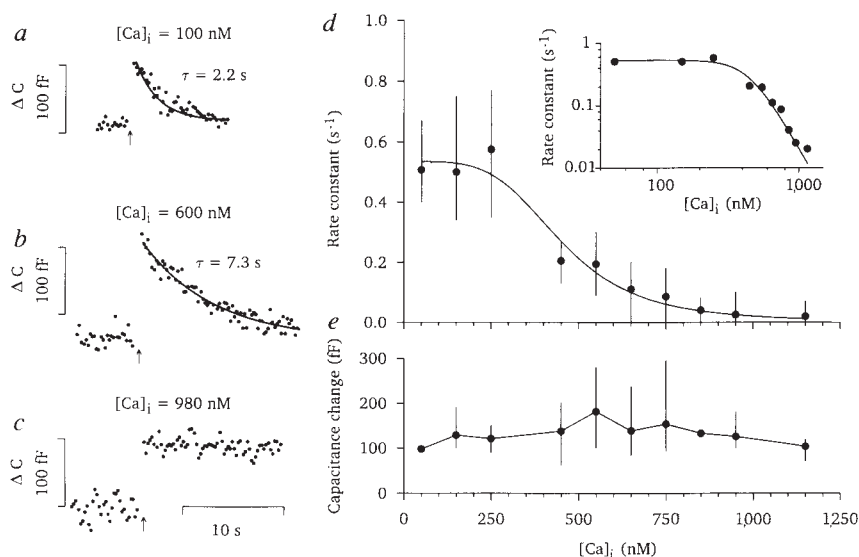
FIG. 1 The onset of endocytosis is delayed during transient elevation of $[Ca^{2+}]_i$ to high levels. **a**, Rapid endocytosis after moderate elevation of $[Ca^{2+}]_i$. The upper trace shows the capacitance change elicited by a single 250-ms pulse from -60 mV to 0 mV (arrow). The inset shows Ca^{2+} current activated by the depolarization. The lower trace shows the resulting increase in $[Ca^{2+}]_i$ (0.5 mM BAPTA, 0.1 mM Fura-2 in pipette). The solid line shows an exponential decay fitted by a least-squares criterion (3 free parameters: baseline capacitance, size of capacitance jump, and time constant, τ). This rapid endocytosis may underlie the high selectivity of membrane retrieval after exocytosis^{17,18} and may be related to secretion by flickering fusion^{13,19}. **b**, Delayed endocytosis after larger elevation of $[Ca^{2+}]_i$. Traces show the capacitance response (upper) and the increase in $[Ca^{2+}]_i$ (lower) elicited by a 1-s depolarization at the arrow (0.4 mM EGTA, 0.2 mM Fura-2 in pipette). The inset shows Ca^{2+} current. $[Ca^{2+}]_i$ reached a higher level than in **a**, and membrane retrieval was delayed after the stimulus. The vertical dashed line indicates the approximate onset of endocytosis, and the horizontal dashed line indicates $[Ca^{2+}]_i$ at the onset of endocytosis (~ 700 nM). Note that, unlike neuroendocrine cells^{12,14,20}, capacitance does not continue to rise after Ca^{2+} -current deactivation. **c**, Elevation of $[Ca^{2+}]_i$ with ionomycin slows the rate of endocytosis. The left trace shows the control capacitance response at resting $[Ca^{2+}]_i$ (250 nM). The right-hand trace shows a response for the same terminal after resting $[Ca^{2+}]_i$ had been elevated to 570 nM by local superfusion with $5 \mu\text{g ml}^{-1}$ ionomycin. Arrows indicate single 250-ms depolarizations. Note the different time-scales for the two traces. For comparison, the dashed line shows the exponential decay from the control response, replotted on the timescale of the right-hand trace. Pipette solution contained (in mM) 5 Cs₂-HEEDTA, 2.62 CaCl₂, 0.2 Fura-2.

METHODS. Bipolar neurons were obtained from goldfish (*Carassius auratus*) retina as described²¹. Whole-terminal patch-clamp recordings were made from acutely isolated synaptic terminals (8–12 μm diameter) at room temperature (20–25 °C). The extracellular oxygenated saline



contained (in mM): NaCl, 115; KCl, 2.6; MgCl₂, 1.6; CaCl₂, 2.5; glucose, 10; HEPES-NaOH, 10; pH 7.3. Patch pipettes were filled with an intracellular solution containing (in mM): Caesium-glucuronate or CsCl, 130; TEA-Cl, 10; MgCl₂, 2; Na₂-ATP, 2; GTP, 0.5; HEPES-CsOH, 25; pH 7.2. Only terminals with series resistance <20 M Ω and with stable leak currents <20 pA were considered. Varying amounts of the Ca^{2+} -buffers BAPTA, EGTA or HEEDTA (*N*-hydroxyethylethylenediaminetriacetic acid) were added to set the level of free Ca^{2+} in the pipette solution as specified in the descriptions of individual experiments. Fura-2 (0.1–0.2 mM) was added to the pipette solution to monitor $[Ca^{2+}]_i$ as described²¹. Membrane capacitance was measured with a sinusoidal voltage-clamp command (800 Hz; 30 mV peak-to-peak; depolarizing pulses were given from the holding potential -60 mV to 0 mV, unless otherwise indicated) and a two-phase lock-in amplifier^{10,11}, or with the automatic capacitance compensation feature of the EPC-9 patch-clamp amplifier¹¹. The two methods gave equivalent results.

FIG. 2 Inhibition of endocytosis is graded with $[Ca^{2+}]_i$. Capacitance responses were elicited by single 250-ms depolarizations (arrows; 180 ms in **c**), and the recovery phase was fitted with an exponential. $[Ca^{2+}]_i$ was buffered as follows (mM): **a**, 0.5 BAPTA, 0.1 Fura-2; **b**, 9 Ca-EGTA, 1 EGTA, 0.2 Fura-2; **c**, 3.1 Ca-HEEDTA, 6.9 HEEDTA, 0.1 Fura-2. The indicated $[Ca^{2+}]_i$ was the average measured value during a one-time-constant window after depolarization. **d**, Relationship between $[Ca^{2+}]_i$ and rate of endocytosis. The rate constant was the reciprocal of the time constant, τ , as shown in **a–c**. Filled circles show means of 4 or 5 experiments (3 and 2 experiments at 50 and 850 nM, respectively). Vertical lines show range of maximum and minimum rates across experiments. Data were pooled in 100-nM windows; only responses recorded from 100 to 200 s after break-in were included. The solid line was drawn according to the Hill equation (see text). The inset shows a double-logarithmic plot to better reveal the fit at high $[Ca^{2+}]_i$. The Hill coefficient was 4, suggesting a cooperative Ca^{2+} -binding protein like calmodulin²². Calmodulin and/or Ca^{2+} bind to proteins associated with synaptic vesicles or their recycling^{23–25}. However, the calmodulin inhibitor W-7 (100 μM) did not affect endocytosis; calmidazolium (10–100 μM) and trifluoperazine (10–100 μM) had nonspecific capacitance effects and could not be tested. Two peptides, calmodulin kinase II inhibitor (50 μM ; RB) and calmodulin-binding domain (60 μM ; Calbiochem), had no effect when tested several diffusion time-constants²⁶ after break-in. Bipolar-



cell terminals are immunoreactive for protein kinase C (ref. 27), which may also regulate endocytosis²⁸. **e**, The average size of the capacitance jump after a 250-ms depolarization did not change consistently with $[Ca^{2+}]_i$, although there was some increase in variability at ~ 600 nM. Same group of terminals as used in **d**.

obtained in seven of nine terminals (average error, 2.8×10^{-5} ; $n=7$). This shows that endocytosis following secretion driven by Ca^{2+} current tracks $[\text{Ca}^{2+}]_i$ on a moment-by-moment basis.

Endocytosis after repetitive stimulation for several seconds is ~ 10 -fold slower than after single brief stimuli^{3,9}. To determine whether Ca^{2+} -dependent inhibition accounts for this slowing, we calculated the expected recovery of capacitance after sustained stimulation in the manner described above. Figure 4a shows that endocytosis after sustained stimulation was slower than predicted (solid line) from the time course of $[\text{Ca}^{2+}]_i$, using the Hill equation of Fig. 2d. This was consistently observed in six experiments; the average mean square error between predicted and actual recovery was 4×10^{-3} for sustained stimuli (compared with 2.8×10^{-5} for brief stimuli). Thus, the slowing of endocytosis after strong stimulation is not due simply to increased elevation of $[\text{Ca}^{2+}]_i$, suggesting that other factors influence the retrieval rate after sustained stimuli. However, Fig. 4b shows that endocytosis following repetitive stimulation was eliminated, as with brief stimuli (Fig. 2c), when $[\text{Ca}^{2+}]_i$ was continuously elevated to $\sim 1 \mu\text{M}$.

A single messenger, internal calcium, regulates both exocytosis and endocytosis in a synaptic terminal. Exocytosis is triggered by high levels of $[\text{Ca}^{2+}]_i$ ($>50 \mu\text{M}$) reached transiently in the near vicinity of release sites during depolarization, implicating a low-affinity Ca^{2+} sensor in exocytosis^{9,15,16}. Activation and deactivation of exocytosis thus occurs rapidly, conferring high temporal fidelity to synaptic transmission. In contrast, endocytosis is inhibited until bulk $[\text{Ca}^{2+}]_i$ in the terminal falls below $\sim 1 \mu\text{M}$. Endocytosis therefore follows exocytosis after a Ca^{2+} -dependent delay. By restricting the supply of membrane for synaptic vesicle recycling, Ca^{2+} -dependent inhibition of endocytosis

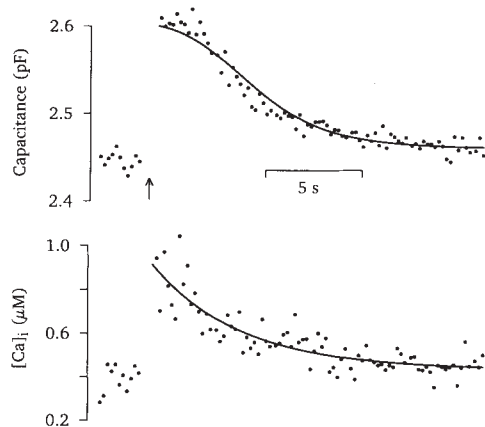


FIG. 3 The return of membrane capacitance to baseline after a depolarizing pulse is described by the change in the rate of membrane retrieval expected from the time course of $[\text{Ca}^{2+}]_i$. The decay of $[\text{Ca}^{2+}]_i$ (bottom trace) was fitted by an exponential function (smooth line) with a time constant, τ_{Ca} , of 4.6 s. The solid line superimposed on the capacitance response (upper trace) was then calculated from the equation $C(t) = 2.46 + 0.14 e^{-t/\tau_{\text{Ca}}}$, where C is capacitance in pF as a function of time, t , after the depolarizing stimulus (750 ms pulse from -60 to 0 mV, given at the arrow; Ca^{2+} current = 200 pA), and α_c is the rate constant of endocytosis. The constants 2.46 pF and 0.14 pF represent the baseline capacitance and the size of the capacitance jump, and α_c was calculated from the Hill equation shown in Fig. 2d: $\alpha_c(t) = \alpha_{\text{max}}(1 - (1/(1 + (K_{1/2}/[\text{Ca}^{2+}]_i)^n)))$, with $\alpha_{\text{max}} = 0.54$, $K_{1/2} = 460$ nM and $n = 4$. $[\text{Ca}^{2+}]_i$ is Ca^{2+} concentration in nM as a function of time, calculated from the exponential relation shown in the bottom trace: $[\text{Ca}^{2+}]_i(t) = 433 + 487 e^{-t/\tau_{\text{Ca}}}$, where $\tau_{\text{Ca}} = 4.6$ s. The deviation of the capacitance data from the predicted line was calculated from the mean square error: $(\sum(C(t) - d(t))^2 \cdot \sum C^2(t) \cdot \sum d^2(t))^{-1/2}$, where $C(t)$ is the predicted and $d(t)$ the actual capacitance at each time point. Pipette solution included (in mM) 10 $\text{Cs}_2\text{-HEEDTA}$, 5.24 CaCl_2 , 0.2 Fura-2.

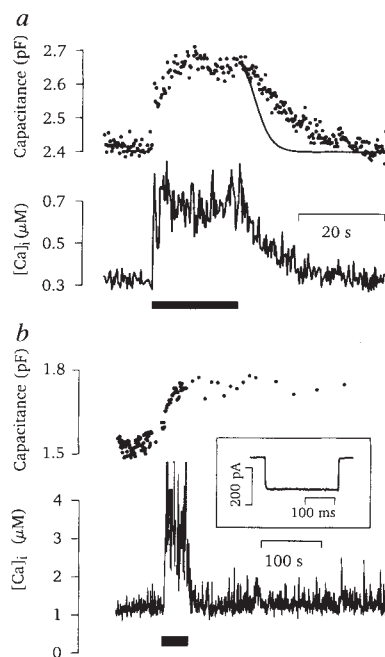


FIG. 4 Membrane retrieval following sustained stimulation is slower than expected from the effect of $[\text{Ca}^{2+}]_i$ on the rate of endocytosis. Previous work has shown two rates of endocytosis, one occurring in seconds and the other in tens of seconds^{3,9}. After sustained stimulation, the slower rate is observed⁹. a, Capacitance response to a series of 250-ms pulses, delivered at 1 pulse per s during the time indicated by the dark bar. Decay of $[\text{Ca}^{2+}]_i$ after termination of stimulation was fitted by an exponential with $\tau = 6.2$ s, and the time course of capacitance was then calculated (solid line) from the Ca^{2+} dependence of the rate constant of endocytosis (Fig. 2d), as described in Fig. 3. b, Elevated $[\text{Ca}^{2+}]_i$ abolishes membrane retrieval following sustained stimulation. A series of 250-ms pulses from -60 mV to -10 mV was presented at 1 pulse per s (dark bar). The inset shows Ca^{2+} current elicited by the first pulse in the series. Resting $[\text{Ca}^{2+}]_i$ was increased by including 9 mM Ca-EGTA and 1 mM free EGTA in the pipette solution, and $[\text{Ca}^{2+}]_i$ was monitored with 0.1 mM Fura-2. Endocytosis following sustained stimulation is thus blocked at high $[\text{Ca}^{2+}]_i$, consistent with previous results⁹, where dialysis of terminals with $52 \mu\text{M}$ $[\text{Ca}^{2+}]_i$ elicited large capacitance increases that plateaued and did not return to baseline. Endocytosis in frog motor nerve terminals stimulated with black-widow spider venom has been reported to require external Ca^{2+} (ref. 29), which conflicts with our conclusion and with ref. 30. However, Hurlbut et al.³¹ recently concluded that this action of external Ca^{2+} is attributable to Ca^{2+} effects on gating of α -latrotoxin channels, rather than a Ca^{2+} requirement for endocytosis.

could retard replenishment of vesicle pools and suppress synaptic output. Responses to prolonged stimuli (Fig. 4), where capacitance rises to a plateau despite continued stimulation, may reflect in part depletion of releasable vesicles in the absence of replenishment via recycling. Because the levels of $[\text{Ca}^{2+}]_i$ necessary to substantially affect endocytosis are reached only with sustained activation of Ca^{2+} current, feedback suppression of vesicle recycling would be most pronounced in terminals that have recently experienced strong stimulation. Ca^{2+} -dependent regulation of endocytosis may thus contribute significantly to activity-dependent changes in synaptic strength in the nervous system. □

Received 28 February; accepted 19 July 1994.

- Heuser, J. E. & Reese, T. S. *J. Cell Biol.* **88**, 564–580 (1980).
- Torri-Tarelli, F., Grohovaz, F., Fesce, R. & Ceccarelli, B. *J. Cell Biol.* **101**, 1386–1399 (1985).
- Miller, T. M. & Heuser, J. E. *J. Cell Biol.* **98**, 685–698 (1984).
- Pysh, J. J. & Wiley, R. G. *J. Cell Biol.* **60**, 365–374 (1974).
- Schaeffer, S. F. & Raviola, E. *J. Cell Biol.* **79**, 802–825 (1978).
- Betz, W. J. & Bewick, G. S. *Science* **255**, 200–203 (1992).
- Katz, B. & Miledi, R. *Proc. R. Soc. B* **161**, 496–503 (1965).
- Mulkey, R. M. & Zucker, R. S. *Nature* **350**, 153–155 (1991).

9. von Gersdorff, H. & Matthews, G. *Nature* **367**, 735–739 (1994).
10. Lindau, M. & Neher, E. *Pflügers Arch. ges. Physiol.* **411**, 137–146 (1988).
11. Neher, E. & Zucker, R. S. *Neuron* **10**, 21–30 (1993).
12. Thomas, P., Suprenant, A. & Almers, W. *Neuron* **5**, 723–733 (1990).
13. Thomas, P., Lee, A. K., Wong, J. G. & Almers, W. *J. Cell Biol.* **124**, 667–675 (1994).
14. Rosenboom, H. & Lindau, M. *Proc. natn. Acad. Sci. U.S.A.* **91**, 5267–5271 (1994).
15. Augustine, G. J., Charlton, M. P. & Smith, S. J. A. *Rev. Neurosci.* **10**, 633–693 (1987).
16. Llinás, R., Sugimori, M. & Silver, R. *Science* **256**, 677–679 (1992).
17. Matteoli, M., Takei, K., Perin, M. S., Südhof, T. C. & De Camilli, P. *J. Cell Biol.* **117**, 849–861 (1992).
18. Nordmann, J. J. & Artault, J.-C. *Neuroscience* **49**, 201–207 (1992).
19. Alvarez de Toledo, G., Fernandez-Chacon, R. & Fernandez, J. M. *Nature* **363**, 554–558 (1993).
20. Chow, R. H. von Rüden, L. & Neher, E. *Nature* **356**, 60–63 (1992).
21. Heidelberger, R. & Matthews, G. *J. Physiol. Lond.* **447**, 235–256 (1992).
22. Sharma, R. K., Mooibroek, M. & Wang, J. H. in *Calmodulin. Molecular Aspects of Cellular Regulation* Vol. 5 (eds Cohen, P. & Klee, C. B.) 265–295 (Elsevier, Amsterdam, 1988).
23. Maycox, P. R., Link, E., Reetz, A., Morris, S. A. & Jahn, R. *J. Cell Biol.* **118**, 1379–1388 (1992).
24. Greengard, P., Valtorta, F., Czernik, A. J. & Benfenati, F. *Science* **259**, 780–785 (1993).
25. Brose, N., Petrenko, A. G., Südhof, T. C. & Jahn, R. *Science* **256**, 1021–1025 (1992).
26. Pusch, M. & Neher, E. *Pflügers Arch. ges. Physiol.* **411**, 204–211 (1988).
27. Suzuki, S. & Kaneko, A. *Vis. Neurosci.* **5**, 223–230 (1990).
28. Robinson, P. J. et al. *Nature* **347**, 163–166 (1993).
29. Ceccarelli, B. & Hurlbut, W. P. *J. Cell Biol.* **87**, 297–303 (1980).
30. Ryan, T. A. et al. *Neuron* **11**, 713–724 (1993).
31. Hurlbut, W.P., Chieregatti, E., Valtorta, F. & Haimann, C. *J. Membrane Biol.* **138**, 91–102 (1994).

ACKNOWLEDGEMENTS. Supported by research grants from the NIH (G.M.) and by a NRSA postdoctoral fellowship (H.V.G.).

Structure and transport mechanism of a high-affinity potassium uptake transporter from higher plants

Daniel P. Schachtman* & Julian I. Schroeder†

Department of Biology and Center for Molecular Genetics, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0116, USA

POTASSIUM is the most abundant cation in higher plants and is crucial for plant nutrition, growth, tropisms, enzyme homeostasis and osmoregulation^{1–4}. K⁺ accumulation can be rate-limiting for agricultural production^{2–4}. K⁺ uptake from soils into roots is largely mediated by high-affinity K⁺ uptake ($K_m \approx 10$ –40 μ M) (refs 1, 2, 5–7). But although K⁺ channels allow low-affinity K⁺ uptake^{8,10}, both the transport mechanism and structure of the high-affinity K⁺ nutrition pathway remain unknown. Here we use expression cloning to isolate a complementary DNA encoding a membrane protein (HKT1) from wheat roots which confers the ability to take up K⁺. The substrate affinity, saturation and cation selectivity of HKT1 correspond to hallmark properties of classical high-affinity K⁺ uptake in plants^{1,2,11}. The transport mechanism of HKT1 uses K⁺-H⁺ co-uptake. Expression of HKT1 is localized to specific root and leaf regions which represent primary sites for K⁺ uptake in plants^{2,3}. HKT1 is important for plant nutrition and could possibly contribute to environmental alkali metal toxicities^{11–13}.

High-affinity K⁺ uptake in roots of higher plants can be induced when plants are deprived of potassium^{6,14}. Therefore to clone a high-affinity K⁺-uptake transporter, a cDNA library was constructed using messenger RNA isolated from 4-day-old wheat roots (*Triticum aestivum*) grown in a medium devoid of K⁺. A mutant yeast strain defective for K⁺ uptake¹⁵ was transformed with purified wheat-root plasmid cDNA. A cDNA, named *HKT1*, was isolated that permitted the mutant yeast strain to grow on media containing 30 μ M K⁺.

The uptake characteristics of yeast cells complemented with the *HKT1* cDNA were analysed using ⁸⁶Rb⁺ as a tracer analogue for K⁺ (ref. 1). In control experiments using the K⁺-uptake-deficient yeast mutant, only very low rates of ⁸⁶Rb⁺ uptake could be detected (Fig. 1a). Conversely, large ⁸⁶Rb⁺-uptake rates were evident in the yeast strain expressing the single wheat-root *HKT1* cDNA (Fig. 1a). Rb⁺-uptake rates in HKT1-expressing yeast could be described by a single Michaelis–Menten function which saturated at ~ 300 μ M Rb⁺ and yielded an apparent equilibrium

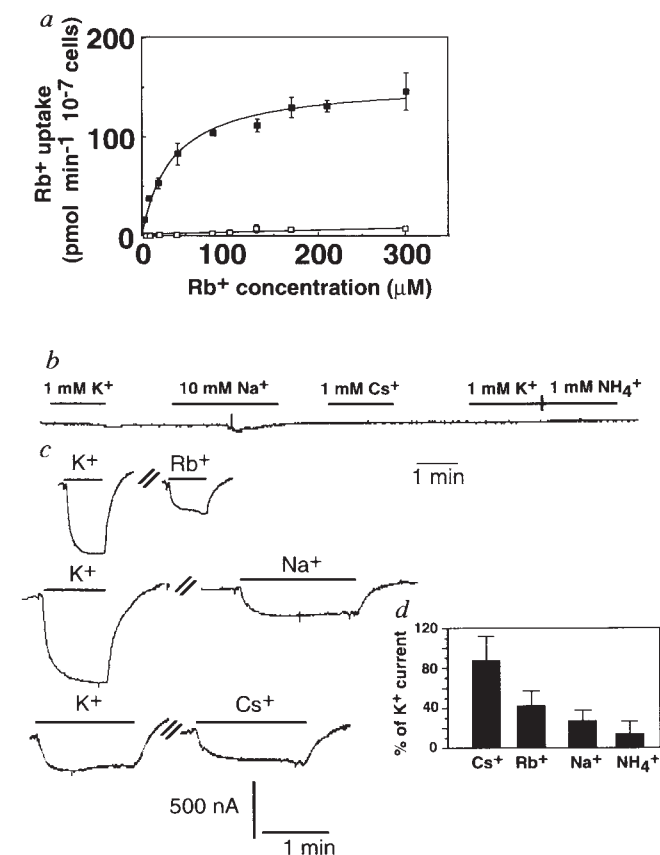


FIG. 1 High-affinity K⁺ uptake and selectivity of HKT1 in yeast (a) and *Xenopus* oocytes (c, d). a, The uptake rate of ⁸⁶Rb⁺ as a function of Rb⁺ concentration in a yeast strain deficient in K⁺ uptake¹⁵ (control: open symbols) and in the same strain expressing HKT1 (filled symbols; Marquardt–Levenberg fit). b, Control water-injected oocytes were exposed to extracellular cations at the indicated times ($V_m = -60$ mV). Current scale as in c. c, Ionic selectivity of HKT1-induced currents in *Xenopus* oocytes. Changes in ionic currents are shown in three HKT1-expressing oocytes in response to extracellular perfusion with (1 mM) K⁺, Rb⁺, Na⁺ and Cs⁺. Perfusion times are indicated above each trace. Currents induced by 1 mM K⁺ are shown for reference. Membrane potential was -80 mV (Na⁺, Cs⁺), -90 mV (Rb⁺). d, HKT1-mediated ionic conductivity ratios ($\pm$ s.d.). Currents carried by each cation were compared to the K⁺ current and measured relative to background currents before cation exposures (as in c).

METHODS. The HKT1-expressing *S. cerevisiae* strain CY162 (ref. 15) was grown overnight then starved of K⁺ for 6 h²². Rb⁺-uptake rates (calculated relative to absorbance at 600 nm) remained linear for >60 min in Ca-MES flux buffer¹⁷. After 40 min, cells were concentrated on 0.45- μ m nylon membrane, washed and analysed in a Beckman LS-230 liquid scintillation counter. Oocyte mRNA injections (15–20 ng), voltage-clamp recordings, data acquisition and analysis were performed as described²³. Oocyte bath solutions contained in mM: 1.8 CaCl₂, 1 MgSO₄, 10 HEPES, pH 7.5, and sorbitol to give an osmolality of 240 mosmol kg⁻¹. K⁺, Na⁺, Cs⁺ and NH₄⁺ were added as glutamate salts, and Rb⁺ as Cl⁻ salt. Background current measurements before and after perfusions in b and c contained 1 mM Tris-glutamate. Perfusion rate was 2 ml min⁻¹ (bath volume ~ 600 μ l).

* Present address: Department of Botany, University of Adelaide, Adelaide, SA 5005, Australia.
† To whom correspondence should be addressed.

dissociation constant (K_m) of $29 \pm 6 \mu\text{M Rb}^+$ ($n=4$; Fig. 1a). These data strongly correlate with high-affinity K^+ accumulation in plant roots, which shows an apparent K_m in the range of $10\text{--}40 \mu\text{M Rb}^+$ and saturation at $\sim 200 \mu\text{M Rb}^+$ (refs 1, 2, 5).

The functional properties of HKT1-mediated high-affinity K^+ uptake were characterized more directly by expression and voltage clamping of HKT1 in *Xenopus* oocytes. In control experiments, when water-injected oocytes were perfused with K^+ or other cations, no changes in ion currents were observed ($n=20$; Fig. 1b). In contrast, large inward currents developed slowly when HKT1-expressing oocytes were exposed to K^+ (Fig. 1c; $n=95$ oocytes, isolated from 16 frogs).

The cation selectivity of HKT1 was characterized by complete replacement of K^+ in the oocyte bathing medium with other alkali metal ions and NH_4^+ . Rubidium-induced currents were $41 \pm 15\%$ as large as K^+ currents ($\pm\text{s.d.}$, $n=15$; Fig. 1c, d). Sodium ($n=19$) and NH_4^+ ions ($n=13$) were transported at low rates relative to the K^+ -uptake rate (Fig. 1c, d). Caesium was transported efficiently into HKT1-expressing oocytes at $\sim 86 \pm 24\%$ the rate of K^+ ions ($n=11$; Fig. 1c, d). A selectivity sequence of $\text{K}^+ > \text{Cs}^+ > \text{Rb}^+ > \text{Na}^+ > \text{NH}_4^+$ was estimated from these results (Fig. 1d).

The biophysical transport mechanism of high-affinity K^+ uptake into higher plant cells remains unknown. Several putative mechanisms have been suggested (for reviews, see refs 3, 4, 16),

including a K^+ -ATPase, a K^+ - H^+ exchanger¹⁶ or a K^+ - H^+ co-uptake transporter¹⁷ energized by the proton-extruding ATPase¹⁸. $^{86}\text{Rb}^+$ -uptake by HKT1-expressing yeast was strongly enhanced by acidification of the extracellular pH from 8 to pH 6–7, suggestive of proton-coupled K^+ uptake (Fig. 2a). At pH 8, Rb^+ uptake by HKT1 was slightly larger than in controls with the non-transformed yeast strain (data not shown).

The transport mechanism of high-affinity K^+ uptake was further characterized by voltage-clamp studies of HKT1-expressing oocytes. The large currents observed in oocytes expressing HKT1 (Fig. 1c) rule out an electroneutral K^+ - H^+ exchanger as the mechanism of high-affinity K^+ accumulation. At depolarized membrane potentials, K^+ -induced currents were outward (Fig. 2b; $n=45$ oocytes) and removal of cations resulted in more positive background currents (Fig. 1c), indicating that exposure to extracellular cations may modify oocyte properties and/or partially enhance HKT1 activity. The mechanism of proton coupling to K^+ uptake was directly investigated by measurements of shifts in the HKT1 conductance (Fig. 2b). In contrast to control oocytes ($n=6$), acidification by one pH unit in the pH range of 8.5 to 5.5 caused positive shifts in the HKT1-mediated conductance by $+24 \pm 5 \text{ mV}$; acidification by two pH units caused a shift of $+58 \pm 7 \text{ mV}$ ($n=7$; Fig. 2b). Furthermore, increases in the extracellular K^+ concentration from 0.5–5 mM caused a zero-current potential shift of $+22 \pm 4 \text{ mV}$ ($n=10$). Absolute zero-current potentials varied, possibly because of other conductance properties. Shifts in the zero-current potential by $+29 \text{ mV}$ per pH unit acidification (Fig. 2b) or per 10-fold increase in K^+ activity would be consistent with a stoichiometry of uptake of ~ 1 proton per K^+ ion by HKT1 (see equation in Fig. 2 legend)¹⁹.

The HKT1 cDNA contains an open reading frame encoding a predicted protein of 534 amino acids (relative molecular mass 58.9 K) (Fig. 3a). HKT1 comprises 10–12 putative membrane-spanning domains²⁰, consistent with its role as a plasma-membrane ion transporter (Fig. 3b). Sequence analysis uncovered no consensus sequences for catalytic ATP-binding sites²¹, suggesting that ATP hydrolysis by HKT1 may not be required as an energy source for HKT1-mediated K^+ uptake. A search of the non-redundant cDNA database showed a weak similarity ($\sim 19\%$ amino-acid identity) to K^+ -uptake transporters from *Saccharomyces* species in some of the hydrophobic domains²². HKT1 hybridizes to one DNA band in both *Arabidopsis thaliana* (Fig. 3c, right) and in a diploid wheat line (Fig. 3c, left) in which the gene was mapped to chromosome 7 (not shown).

The expression pattern of HKT1 RNA was determined by *in situ* hybridization to sections of wheat tissue. HKT1 expression was primarily localized to cortical cells in roots (Fig. 4a–c) and to cell layers bordering the vascular tissue in leaves (Fig. 4e–h). HKT1 expression in roots and leaves is consistent with data indicating high-affinity K^+ uptake in several plant tissues². Expression of HKT1 in the root cortex (Fig. 4a–c) supports indications that cortical cells have a large K^+ -uptake activity during K^+ nutrition from soils^{5,7}. Transport of K^+ from roots to all leaf cells requires K^+ uptake by cells surrounding the vascular tissues in leaves. The expression pattern of HKT1 to several cell layers surrounding the leaf vascular tissue (Fig. 4e–h) suggests that HKT1 may be essential for nutrient transfer of K^+ into leaves.

The functional properties of HKT1 provide direct evidence for the classical hypothesis of Epstein *et al.* that higher plants possess a high-affinity K^+ -accumulation pathway which is distinct from low-affinity K^+ -uptake transporters¹. Inward-rectifying K^+ channels provide one molecular pathway for low-affinity, cell-specific, K^+ uptake^{4,8–10,15,23}. The electrogenic nature of HKT1-mediated transport (Fig. 1c), the lack of a catalytic ATP-binding domain (Fig. 3a), as well as the pH dependency of Rb^+ -uptake (Fig. 2a) and of HKT1-associated currents (Fig. 2b), together suggest that K^+ - H^+ co-uptake is the biophysical transport mechanism of high-affinity K^+ uptake into higher plant

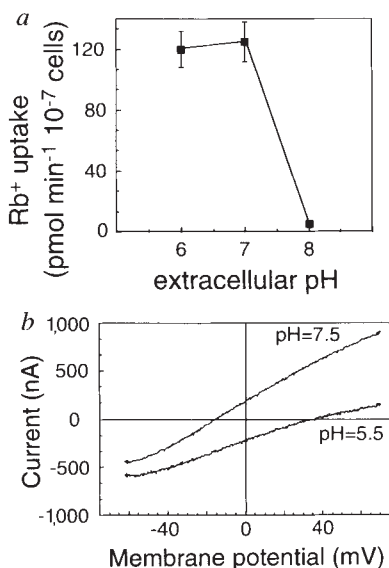
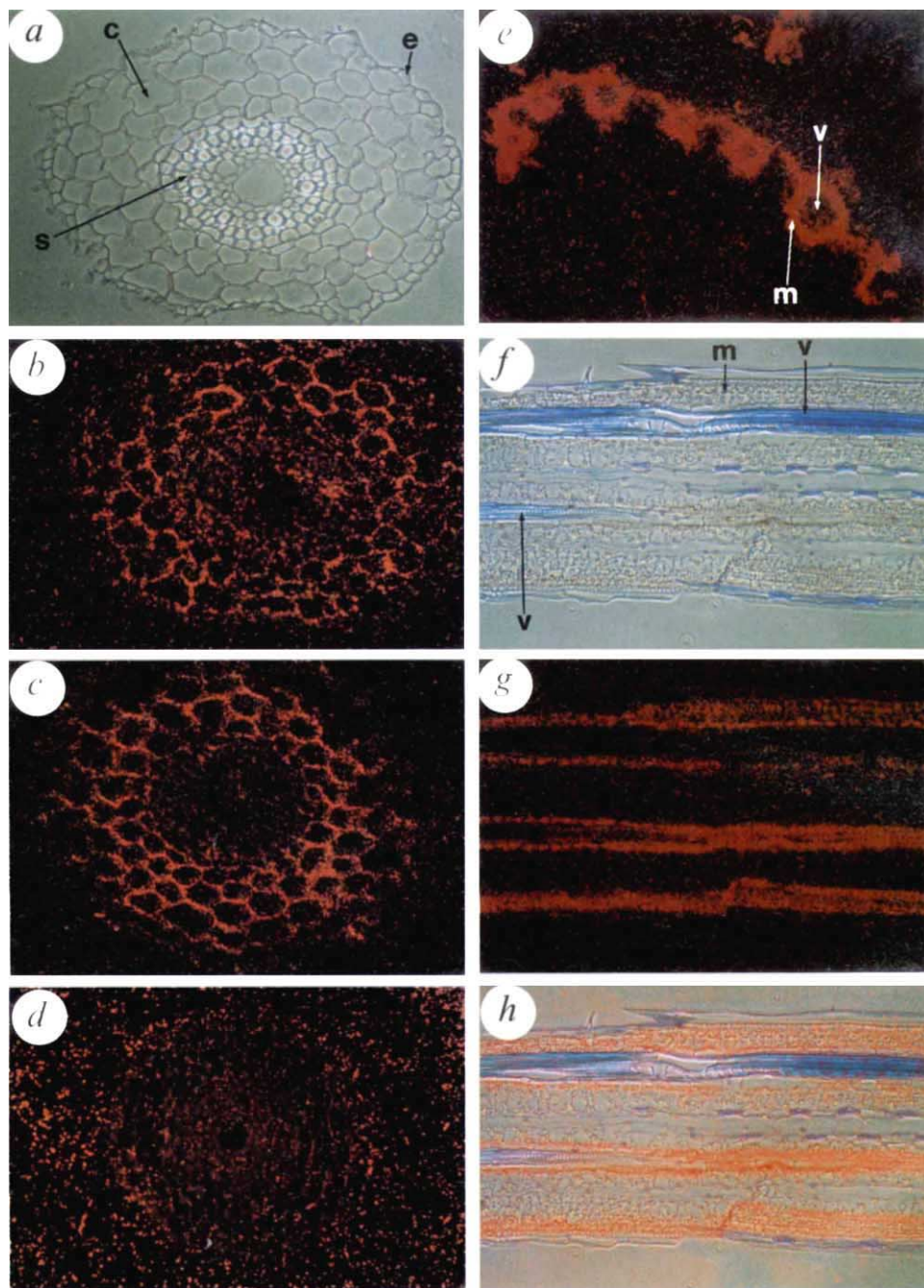


FIG. 2 Uptake rate (a) and conductance shift (b) of HKT1 depends on the extracellular pH. a, Extracellular acidification from pH 8 to pH 7 and 6 caused a dramatic enhancement of Rb^+ uptake in HKT1-expressing yeast. Further acidification to pH 5.0 and 4.0 reduced the rate of Rb^+ uptake when compared to pH 7.0. b, Shift of the HKT1-associated conductance in *Xenopus* oocytes in response to extracellular acidification. In several experiments, upon acidification HKT1-mediated currents did not reverse with respect to oocyte background current in 1 mM Tris-glutamate.

METHODS. In Rb^+ -uptake experiments the Ca-MES buffer¹⁷ was adjusted to pH 7 and 8 with 10 mM HEPES in place of MES (Fig. 1 methods; $100 \mu\text{M Rb}^+$). Oocyte background currents (in 1 mM Tris-glutamate) were subtracted from currents induced by exposure to 1 mM K^+ glutamate, pH 7.5 and 5.5 (solutions as for Fig. 1) to obtain the illustrated current-voltage relationships in b and to determine conductance shifts of HKT1-associated currents. Voltage ramps (duration 3 s) were applied to oocytes from a holding potential of -60 mV and data were low-pass filtered at 20 Hz. Zero-current potential (V_0) shifts were calculated using the equation¹⁹ derived for co-uptake of one H^+ per K^+ ion: $V_0 = 58/2 \text{ mV} (\Delta\text{pH} + \log ([\text{K}^+]_{\text{ext}}/[\text{K}^+]_{\text{cyt}}))$, where $[\text{K}^+]_{\text{ext}}$, $[\text{K}^+]_{\text{cyt}}$, and ΔpH are the external and cytoplasmic K^+ activities and the pH gradient ($\text{pH}_{\text{cyt}} - \text{pH}_{\text{ext}}$), respectively.

FIG. 4 *In situ* hybridization of *HKT1* RNA illustrates the expression pattern in roots and leaves of 4-day-old wheat seedlings grown in K^+ -free medium. Root tissue is shown in cross sections (a–d) and leaf tissue is shown in cross (e) and longitudinal sections (f–h). a, Epidermal (e), cortical (c) and stele (s) regions of the root are clearly visible. b, The hybridization pattern (in red) of the *HKT1* probe shows high levels of *HKT1* expression in root cortical cells and background levels in the stele. Same cross-section as in a. c, Expression pattern of *HKT1* in a different root cross-section. d, Control *in situ* hybridization. e, The *HKT1* probe hybridizes primarily to several layers of mesophyll cells (m) surrounding the vascular tissue (v) from a leaf cross-section. f, The vascular tissue (v) stains darker blue than the mesophyll (m) in longitudinal leaf sections. g, h, Expression of *HKT1* is high in several mesophyll cell layers surrounding the vascular tissue. Same section as in f. Bright-field images in a, f; dark-field images with silver grains enhanced by a red filter in b–e, g; and a double exposure with a dark field (red) superimposed on a bright field image in h.

METHODS. *HKT1* cDNA was used as a probe after 158 nucleotides including the poly(A)⁺ tail were removed from the C terminus. ³⁵S-labelled antisense mRNA was synthesized from *HKT1* (G + C content, 47%). Sections were 8–10 μ m thick. Hybridization conditions were as described³⁰. Slides were exposed for 1–2 weeks. A flower-specific *Arabidopsis* gene (*CAL*) was used as a control (G + C content, 46%; S. Kempin, B. Savidge and M. Yanofsky, unpublished).



transport and of related environmental stresses such as K^+ deficiency and alkali metal toxicity in higher plants. □

Received 5 April; accepted 18 July 1994.

- Epstein, E., Rains, D. W. & Elzam, O. E. *Proc. natn. Acad. Sci. U.S.A.* **49**, 684–692 (1963).
- Epstein, E. *Nature* **212**, 1324–1327 (1966).
- Kochian, L. V. & Lucas, W. J. *Adv. Botany Res.* **15**, 93–177 (1988).
- Schroeder, J. I., Ward, J. M. & Gassmann, W. A. *Rev. Biophys. biomolec. Struct.* **23**, 441–471 (1994).
- Newman, I. A., Kochian, L. V., Grusak, M. A. & Lucas, W. J. *Plant Physiol.* **84**, 1177–1184 (1987).
- Fernando, M., Mehrotra, J. & Glass, A. D. M. *Plant Physiol.* **100**, 1269–1276 (1992).
- Maathuis, F. J. M. & Sanders, D. *Planta* **191**, 302–307 (1993).
- Schroeder, J. I., Raschke, K. & Neher, E. *Proc. natn. Acad. Sci. U.S.A.* **84**, 4108–4112 (1987).
- Schroeder, J. I. & Fang, H. H. *Proc. natn. Acad. Sci. U.S.A.* **88**, 11583–11587 (1991).
- Gassmann, W. & Schroeder, J. I. *Plant Physiol.* (in the press).
- Rains, D. W. & Epstein, E. *Science* **148**, 1611 (1965).
- Welch, R. M. & Epstein, E. *Proc. natn. Acad. Sci. U.S.A.* **61**, 447–453 (1968).
- Sheahan, J. J., Ribeiro-Neto, L. & Sussman, M. R. *Plant J.* **3**, 647–656 (1993).
- Drew, M. C., Saker, L. R., Barber, S. A. & Jenkins, W. *Planta* **160**, 490–499 (1984).
- Anderson, J. A., Huprikar, S. S., Kochian, L. V., Lucas, W. J. & Gaber, R. F. *Proc. natn. Acad. Sci. U.S.A.* **89**, 3736–3740 (1992).

- Poole, R. J. *A. Rev. Plant Physiol.* **29**, 437–460 (1978).
- Rodríguez-Navarro, A., Blatt, M. R. & Slayman, C. L. *J. gen. Physiol.* **87**, 649–674 (1986).
- Spanswick, R. M. *A. Rev. Plant Physiol.* **32**, 267–289 (1981).
- Lester, H. A., Mager, S., Quick, M. W. & Corey, J. L. *A. Rev. Pharmac. Toxicol.* **34**, 219–249 (1994).
- Kyte, J. & Doolittle, R. F. *J. molec. Biol.* **157**, 105–132 (1984).
- Higgins, C. F. *et al. Nature* **323**, 448–450 (1986).
- Ko, C. H. & Gaber, R. F. *Molec. cell. Biol.* **8**, 4266–4273 (1991).
- Schachtman, D. P., Schroeder, J. I., Lucas, W. J., Anderson, J. A. & Gaber, R. F. *Science* **258**, 1654–1658 (1992).
- Maathuis, F. J. M. & Sanders, D. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
- Greenway, H. & Munns, R. A. *Rev. Plant Physiol.* **31**, 149–190 (1980).
- Rains, D. W. & Epstein, E. *Plant Physiol.* **42**, 314–318 (1967).
- Kozak, M. *Nucleic Acids Res.* **12**, 857–872 (1984).
- Schachtman, D. P., Munns, R. & Whitecross, M. I. *Crop. Sci.* **31**, 992–997 (1991).
- Frommer, W. B., Hummel, S. & Riesmeier, J. W. *Proc. natn. Acad. Sci. U.S.A.* **90**, 5944–5948 (1993).
- Drews, G. N., Bowman, J. L. & Meyerowitz, E. M. *Cell* **65**, 991–1002 (1991).

ACKNOWLEDGEMENTS. We thank R. F. Gaber for the CY162 yeast strain; M. Yanofsky, C. Gustafson-Brown and B. Savidge for help with *in situ* hybridizations; F. Rubio for help with some membrane potential recordings; W. Frommer for discussions; E. Lagudah and S. Chandramohan for mapping *HKT1*; and W. Gassmann, N. Spitzer, J. Ward, J. Kyte, M. Yanofsky, D. Weigel, J. Umbach, C. Gunderson and H. Lester for comments on the manuscript. This work was supported by an NSF Presidential Young Investigator Award and grants from the NSF and the Department of Energy (J.I.S.) and by an NSF Plant Biology Postdoctoral Fellowship (D.P.S.).

Interactions of a small RNA with antibiotic and RNA ligands of the 30S subunit

Prakash Purohit & Seth Stern

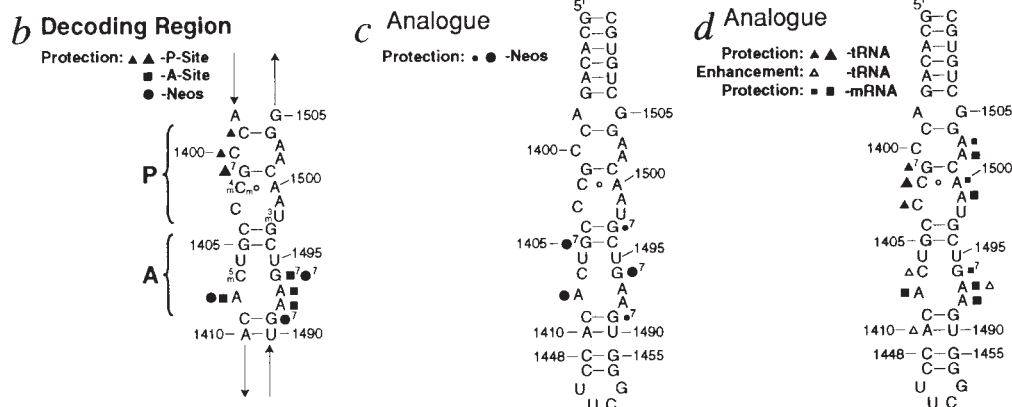
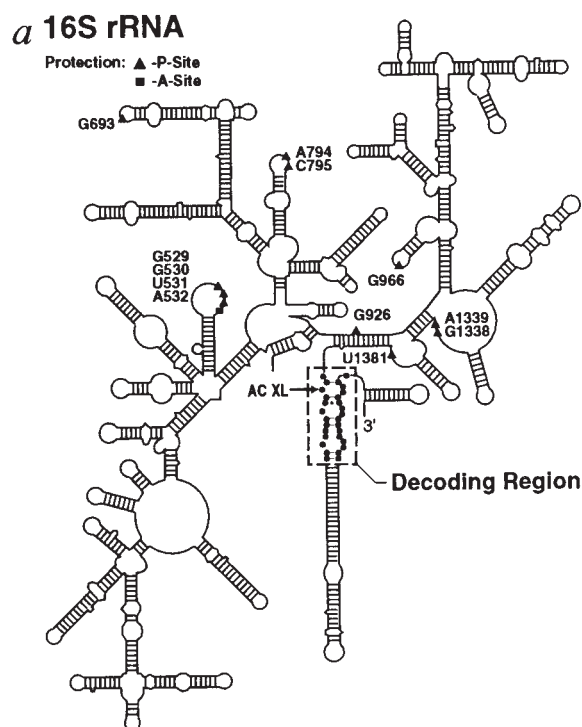
Program in Molecular Medicine, UMASS Medical Center, 373 Plantation Street, Worcester, Massachusetts 01605, USA

It is now generally accepted that 16S and 23S ribosomal RNA play important roles in the decoding and peptidyl transferase activities of ribosomes^{1,2}. Despite their complex structures and numerous associated proteins it is possible that small domains of these rRNAs can fold and function autonomously, particularly those that appear devoid of protein interactions³. One candidate for such a domain is the decoding region, located near the 3' end of 16S rRNA (Fig. 1a, b). Consistent with this hypothesis, aminoglycoside antibiotics that interact with the decoding region in 30S subunits interact with other RNAs in the absence of proteins⁴⁻⁷. In addition, certain activities of self-splicing introns, at least superficially, resemble translational decoding^{8,9}. We report here that an oligoribonucleotide analogue of the decoding region interacts with both antibiotic and RNA ligands of the 30S subunit in a manner that correlates with normal subunit function. The activities of the

decoding region analogue suggest that the intimidating structural complexity of the ribosome can be, to some degree, circumvented.

We designed an oligoribonucleotide analogue of the decoding region shown in Fig. 1c, d, and asked first if it could bind aminoglycoside antibiotics such as neomycin, which protect N1 of A1,408 and N7 of G1,491 and G1,494 within the A-site subdomain of the decoding region^{10,11}; similar interactions with the analogue would indicate its proper folding and functional potential. Conversely, other antibiotics, such as streptomycin, tetracycline and erythromycin, interact with other segments of 16S rRNA or 23S rRNA¹⁰⁻¹² and should not, therefore, interact with the analogue. The analogue was probed with dimethyl sulphate¹³ (DMS) either alone, as naked RNA, or in the presence of increasing concentrations of neomycin, paromomycin, hygromycin, streptomycin, tetracycline and erythromycin. Neomycin (Fig. 2a, lanes 7-10) and the closely related aminoglycoside paromomycin (lanes 11-14) strongly protected N1 of A1,408 at concentrations of 10 and 1 μ M, respectively. In addition, hygromycin (lanes 15-18), a structurally dissimilar aminoglycoside, weakly enhanced the reactivity of N1 of A1,408. In contrast, streptomycin (lanes 19-22), tetracycline (lanes 23-26) and erythromycin (lanes 27-30), antibiotics that do not interact with the decoding region in 16S rRNA, did not detectably interact with the analogue. In addition to N1 of A1,408, N7 of G1,405

FIG. 1 16S rRNA, the decoding region, oligoribonucleotide analogues and probing data. a, Schematic diagram of 16S rRNA showing nucleotides outside the decoding region (boxed) implicated in tRNA-30S subunit interactions at the A- and P-sites^{1,15,16}. The position of a zero-length crosslink between the tRNA anticodon loop and C1,400 in the decoding region (AC XL) is indicated²¹. b, The decoding region, showing nucleotides protected from DMS modification (at N1 of A, N3 of C, and N7 of G) by neomycin-like aminoglycoside antibiotics^{10,11} (Neos) and A- and P-site tRNA^{15,16}. The positions of the A- and P-site decoding region subdomains referred to in the text, and the identities of post-transcriptionally methylated nucleotides²², are indicated. The decoding region is drawn with the tertiary base pairs of Gutell and co-workers^{23,24}. C1,399-G1,504, G1,401-C1,501, C1,402-A1,500, C1,404-G1,497 and G1,405-C1,496. c, The decoding region analogue with nucleotides protected from DMS modification by neomycin and paromomycin (Neos) indicated. The analogue contains 16S rRNA decoding region sequences from A1,398 to A1,410 on the proximal side and U1,490 to G1,505 on the distal side. The distal end of the analogue terminates in the same stem-tetraloop found at the end of the penultimate stem of 16S rRNA (nucleotides 1,448-1,455). The heterologous clamping helix is shown at the top. Weak and strong protections are indicated by small and large symbols, respectively. d, The decoding region analogue with nucleotides protected from DMS modification, or enhanced in reactivity towards DMS, in response to poly(U) mRNA and tRNA anticodon stem-loop T7 transcripts.



and G1,494 were strongly protected by neomycin and paromomycin (Fig. 2b), whereas N7 of G1,491 and G1,497 were weakly protected. These data are summarized in Fig. 1c. Because these results closely match those previously obtained with antibiotics and ribosomes¹⁰⁻¹², we concluded that it is likely that the A-site subdomain of the analogue folds into a biologically relevant conformation.

We asked next if the analogue could function similarly to the decoding region of 16S rRNA in the ribosome by interacting with transfer RNA and/or messenger RNA. In the experiment shown in Fig. 3 increasing concentrations of a minimal tRNA for these purposes¹⁴, a T7 transcript of the anticodon stem-loop of *Escherichia coli* phenylalanine-specific tRNA (tRNA^{phe}), were incubated either alone (lanes 7-10 and 17-20) or in the presence of cognate poly(U) message (lanes 11-15 and 21-25). Inspection of Fig. 3 shows that the anticodon stem-loop transcript protected N3 of C1,402 and C1,403 (compare lane 6 with lanes 7-10) and N7 of G1,401 (compare lane 16 with lanes 17-20), whereas poly(U) protected N1 of A1,408, A1,492, A1,493, A1,499 and A1,502 strongly, N1 of A1,500, A1,503 weakly, and N7 of G1,494 weakly (compare lane 6 with lanes 11-15 and lane 16 with lanes 21-25). In addition, the reactivities of N1 of A1,410 and A1,493, and N3 of C1,407 were enhanced by the stem-loop (compare lane 6 with lanes 7-10), effects which were attenuated in the presence of poly(U) (compare lanes 7-10 with lanes 12-15). Comparison of the present results with

data from previous studies with ribosomes^{15,16} shows that the group of nucleotides protected by the tRNA anticodon stem-loop transcript in the analogue (G1,401, C1,402 and C1,403) are closely associated with, and in fact overlap, those protected by P-site-bound tRNA in the presence or absence of message (C1,399, C1,400 and G1,401). Furthermore, within the A-site subdomain of the analogue nucleotides protected by mRNA (A1,408, A1,492-G1,494) are identical to those associated with the mRNA-dependent interaction of tRNA with the A-site of ribosomes^{15,16}. The other nucleotides, protected by mRNA (A1,499, A1,500, A1,502 and A1,503), have not been previously identified with either A- or P-site function in chemical probing experiments.

To help determine whether poly(U) mRNA interactions were mediated by Watson-Crick base pairs with A residues in the analogue we tested poly(C), an mRNA that cannot form Watson-Crick base pairs with A residues, and polydeoxy(U), a 2'-deoxy-form of poly(U) with similar Watson-Crick base-pairing potential. The experiment of Fig. 4a showed that while poly(U) protected A1,408, A1,492 and A1,493 strongly as expected (compare lanes 6-8 with lanes 9-11), polydeoxy(U) failed to interact with any nucleotides in the analogue (compare lanes 6-8 with lanes 12-14). The interactions of poly(C) were more complex, as the reactivities of N1 of A1,410 and N3 of C1,402 and C1,403 were enhanced (compare lanes 6-8 with lanes 15-17). However, with the exception of A1,492, poly(C) protected the same

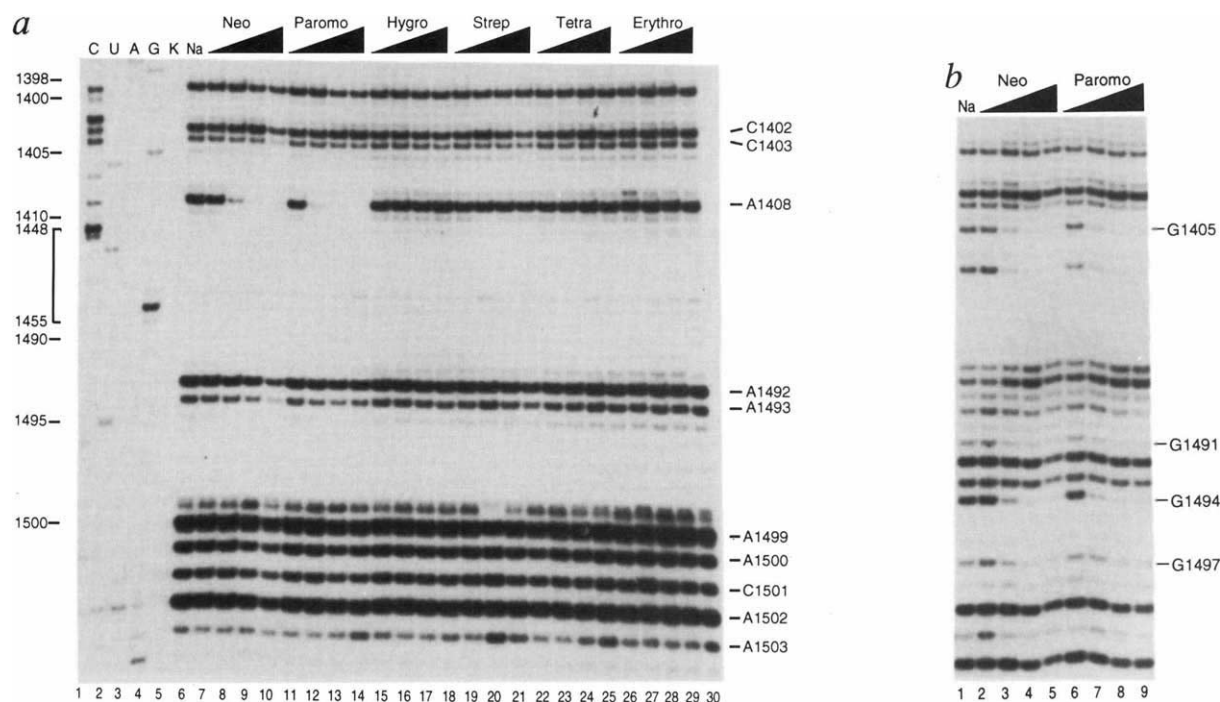


FIG. 2 Decoding region analogue-antibiotic interactions. a, DMS probing reactions with the analogue alone (lane 6) and in the presence of 0.1, 1, 10 and 100 μ M neomycin (lanes 7-10), paromomycin (lanes 11-14), hygromycin (lanes 15-18), streptomycin (lanes 19-22), tetracycline (lanes 23-26) and erythromycin (lanes 27-30). Dideoxy sequencing reactions (lanes 1-4) and a control extension reaction with unmodified RNA (K, lane 5) are to the left, and bands corresponding to nucleotides discussed in the text are indicated to the right. b, DMS/N7 probing reactions with the analogue alone (lane 1) and in the presence of 0.1, 1, 10 and 100 μ M neomycin (lanes 2-5), paromomycin (lanes 6-9). METHODS. Decoding region analogue RNA was transcribed with T7 RNA polymerase²⁵ from a linearized pGEM3 plasmid⁷ containing analogue sequences flanked by *Eco*RI and *Bam*HI restriction sites, and a 15mer reverse primer annealing site immediately 3' of the *Bam*HI site. Gel purified RNA was concentrated by ethanol precipitation, heated to 80 °C

for 1 min, and immediately placed at 37 °C for 5 min. Binding reactions (12.5 μ l) containing 125 ng analogue RNA and antibiotics in 80 mM K-HEPES (pH 7.9), 50 mM NH_4Cl , and 5% PEG buffer were annealed at 37 °C for 15 min and incubated on ice for 1 h. DMS (1 μ l of 1:5 in ethanol) was added, and modification reactions incubated for 40 min on ice. Reactions were stopped with DMS Stop²⁶ and the RNA was purified by ethanol precipitation. DMS/N7 reactions were done according to standard protocols²⁶ except that lyophilization steps were replaced by acid-phenol extraction and ethanol precipitation. For primer extension, 10 ng analogue RNA was annealed to 0.75 ng end-labelled primer and extended with 10-15 U MoMuLV reverse transcriptase for 1 h. Reactions were stopped by ethanol precipitation, pellets resuspended in 10 μ l 8 M urea, 0.05 $\times$ TBE loading buffer, and 2 μ l was loaded onto 8%, 19:1 acrylamide:bisacrylamide, 0.5 $\times$ TBE sequencing gels.

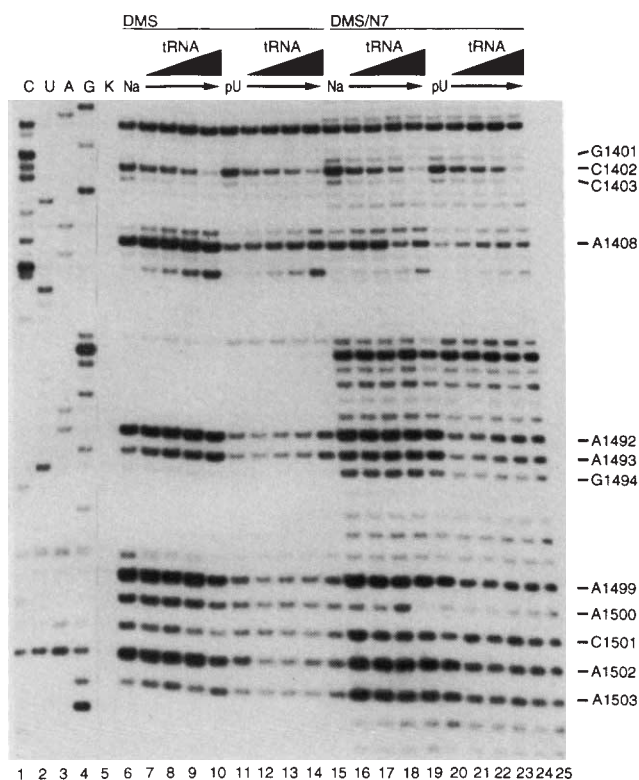


FIG. 3 Interactions of tRNA anticodon stem-loop transcript and poly(U) with the decoding region analogue. DMS and DMS/N7 probing reactions with naked analogue (lanes 6 and 16), and 3, 6, 12 and 24 μ M anticodon stem-loop transcript (lanes 7–10 and 12–15). The same reactions were repeated in the presence of 4 μ g poly(U) (lanes 11–15 and 21–25).

METHODS. Binding reactions were as described for Fig. 2 except that binding buffer contained 200 mM NH_4Cl and 80 mM MgCl_2 . *E. coli* tRNA^{phe} anticodon stem-loop (sequence GGGGAUUGAAAAUCCCC) was transcribed with T7 RNA polymerase, gel purified, concentrated by ethanol precipitation, and annealed with a brief heat step (80° for 1 min followed by 10 min on ice).

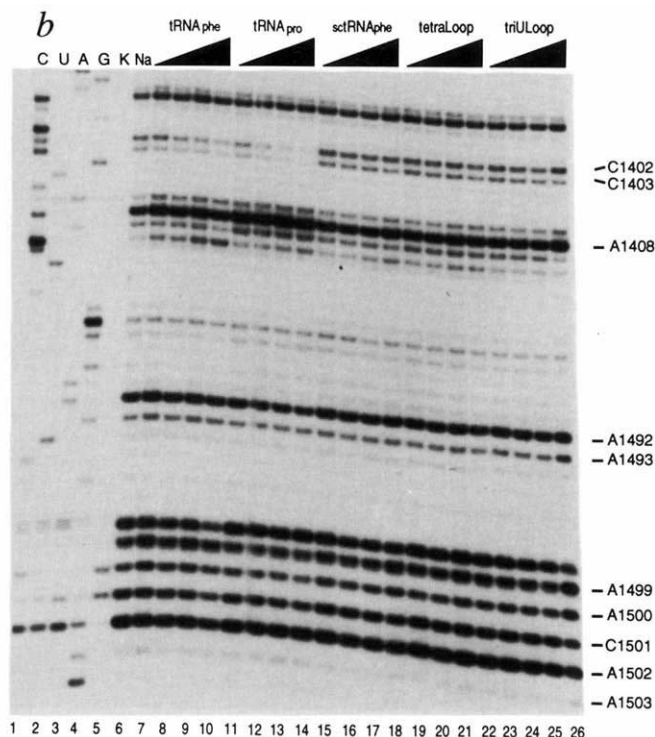
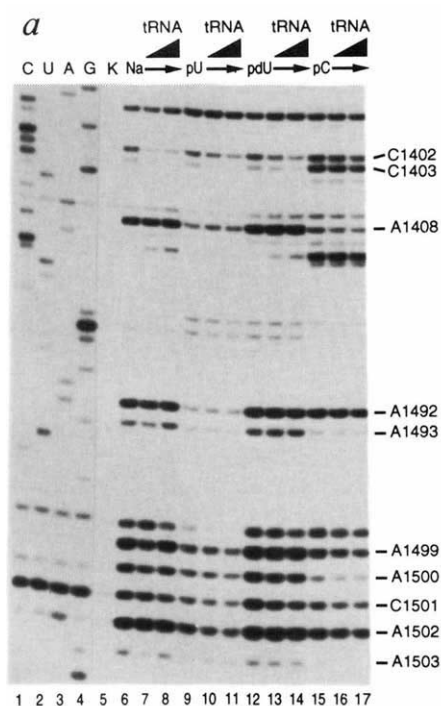


FIG. 4 Selectivity of decoding region analogue interactions. *a*, Selectivity of mRNA interactions. DMS probing reactions with naked analogue (lane 6), 3 μ M (lane 7), and 6 μ M (lane 8) anticodon stem-loop transcript were repeated in the presence of 2 μ g poly(U) (lanes 9–11), polydeoxy(U) (lanes 12–14), and poly(C) (lanes 15–17). Nucleotides protected by tRNA anticodon stem-loop or mRNA are indicated to the right. *b*, Selectivity of the P-site subdomain for stem-loop structures. DMS probing reactions with naked analogue (lane 6), and 1.5, 3, 6 and 12 μ M *E. coli* tRNA^{phe} anticodon stem-loop transcript (lanes 7–10), tRNA^{pro} anticodon stem-loop transcript (lanes 11–14), scrambled

tRNA^{phe} (sctRNA^{phe}, lanes 15–18), tetraloop element (lanes 19–22), and triULoop element (lanes 23–26). Nucleotides protected by the anticodon stem-loops are indicated to the right.

METHODS. Reactions were as described for Fig. 3. tRNA^{pro} anticodon stem-loop (sequence GGUCAUCUUGGGUGAUGACC), scrambled tRNA^{phe} (sequence GGGAGCGUCAUCACAU), tetraloop element (sequence GGGACUUCGGUCCC), and triULoop element (sequence GGCGCUUUGCGCC) were transcribed and treated as described for *E. coli* tRNA^{phe} anticodon stem-loop.

nucleotides as poly(U), including A1,408, A1,493, A1,499, A1,500, A1,502 and A1,503. Thus, despite the apparent disruption of analogue structure induced by poly(C), its protection pattern was virtually identical to that of poly(U). Considered together, these results suggest that the interactions of mRNA with the decoding region analogue are not mediated simply by Watson-Crick base pairing.

C1,402 and C1,403 are unreactive in 16S rRNA¹⁷; their reactivity in the analogue and their protection by the tRNA anticodon stem-loop transcript suggest that the structure of the P-site subdomain of the analogue differs from that of ribosomes. To understand better the discriminatory capacity of the P-site subdomain of the analogue we asked whether it differentiates between closely related RNA structures. In the experiment shown in Fig. 4b a second (sequence-divergent) proline-specific tRNA anticodon stem-loop (tRNA_{Pro}), a scrambled-sequence *E. coli* tRNA_{Phe} with little potential for secondary structure formation (scRNA_{Phe}), a highly stable tetraloop stem-loop element^{18,19} (tetraloop), and another, smaller, stem-loop element with a loop composed of only three U residues²⁰ (triUloop) were each titrated into a binding reaction with the analogue. Only the tRNA_{Phe} and tRNA_{Pro} anticodon stem-loop transcripts interacted, each strongly protecting C1,402 and C1,403 (Fig. 4b). The failure of the scrambled tRNA_{Phe} and other stem-loop elements to interact detectably supports the hypothesis that, despite its altered conformation, the P-site subdomain interacts selectively with RNA ligands. These results suggest, therefore, that the P-site subdomain of the analogue has discriminatory capacity that may be related to ribosomal P-site function.

It would be remarkable indeed if omission of 1,500 nucleotides, all nucleotide modifications, and the 21 small subunit proteins had no effect on the structure and functional activities of the decoding region of 16S rRNA. In light of these omissions, the interactions between the decoding region analogue and 30S subunit ligands that correlate with normal 30S subunit activities are, perhaps, surprising. Positive correlations are probably best exemplified by the interactions of the antibiotics with the A-site subdomain where a nearly perfect correspondence between the characteristics of the decoding region in ribosomes and the analogue was observed with a wide variety of compounds. The interactions of the A-site subdomain with mRNA also showed interesting correlations with previous studies. Precisely the same set of A-site subdomain nucleotides (A1,408, A1,492–G1,494) are protected in the analogue and in ribosomes, where the strict mRNA-dependence of these interactions has made it impossible to determine whether mRNA, tRNA, or both, directly interact with these residues¹⁵. Our results suggest that mRNA protects these nucleotides in ribosomes, and that the analogue simply does not require the simultaneous presence of tRNA to potentiate the interactions. In contrast, interpretation of the significance of the interactions of the P-site subdomain of the analogue with mRNA and tRNA stem-loop transcripts is more complex because, although there are clear similarities with the behaviour of ribosomes, differences between the two systems were also detected. Nonetheless, the parallels between the activities of 30S subunits and the decoding region analogue which we have characterized show that antibiotic, mRNA and, perhaps, tRNA anticodon stem-loop binding activities are intrinsic functions of the decoding region RNA. Given the comparative structural simplicity of the decoding region oligoribonucleotide analogue, it is likely that this and similar RNA constructs will help studies of basic mRNA decoding mechanisms in the future. □

6. von Ahsen, U. & Noller, H. F. *Science* **260**, 1500–1503 (1993).
7. Zapp, M. L., Stern, S. & Green, M. R. *Cell* **74**, 969–978 (1993).
8. Noller, H. F. *Nature* **353**, 302–303 (1991).
9. Schroeder, R., Streicher, B. & Wank, H. *Science* **260**, 1443–1444 (1993).
10. Moazed, D. & Noller, H. F. *Nature* **327**, 389–394 (1987).
11. Woodcock, J., Moazed, D., Shannon, M., Davies, J. & Noller, H. F. *EMBO J.* **10**, 3099–3103 (1991).
12. Moazed, D. & Noller, H. F. *Biochimie* **69**, 879–884 (1987).
13. Stern, S., Moazed, D. & Noller, H. F. *Meth. Enzym.* **164**, 481–489 (1988).
14. Rose, S. J., Lowary, P. T. & Uhlenbeck, O. C. *J. molec. Biol.* **167**, 103–117 (1983).
15. Moazed, D. & Noller, H. F. *Cell* **47**, 985–994 (1986).
16. Moazed, D. & Noller, H. F. *J. molec. Biol.* **211**, 135–145 (1990).
17. Moazed, D., Stern, S. & Noller, H. F. *J. molec. Biol.* **187**, 399–416 (1986).
18. Tuerk, C. et al. *Proc. natn. Acad. Sci. U.S.A.* **85**, 1364–1368 (1988).
19. Woese, C. R., Winker, S. & Gutell, R. R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 8467–8471 (1990).
20. Davis, P. W., Thurmes, W. & Tinoco, I. Jr. *Nucleic Acids Res.* **21**, 537–545 (1993).
21. Prince, J. B., Taylor, B. H., Thurlow, D. L., Ofengand, J. & Zimmerman, R. *Proc. natn. Acad. Sci. U.S.A.* **79**, 5450–5454 (1982).
22. Noller, H. F. *A. Rev. Biochem.* **53**, 119–162 (1984).
23. Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. *Prog. Nucleic Acids Res. molec. Biol.* **32**, 153–216 (1985).
24. Gutell, R. R. in *The Translational Apparatus* (eds Nierhaus, K. H. et al.) 477–488 (Plenum, New York, 1993).
25. Milligan, J. F. & Uhlenbeck, O. C. *Meth. Enzym.* **180**, 51–62 (1989).
26. Peattie, D. A. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **77**, 4679–4682 (1980).

ACKNOWLEDGEMENTS. We thank H. Noller, D. Moazed and T. Powers for their advice, comments, and suggestions for improvements to the manuscript and O. Uhlenbeck, M. Yarus, M. Ares and M. Czech for their comments and suggestions.

Crystal structure of the extracellular region of human tissue factor

K. Harlos*, D. M. A. Martin†§, D. P. O'Brien†§, E. Y. Jones*, D. I. Stuart*, I. Polikarpov†, A. Miller†§, E. G. D. Tuddenham†§ & C. W. G. Boys†||

* Laboratory of Molecular Biophysics and Oxford Centre for Molecular Sciences, Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK

† Haemostasis Research Group, MRC Clinical Research Centre, Watford Road, Harrow, Middlesex HA1 3UJ, UK

‡ Department of Biochemistry, The Medical School, Hugh Robson Building, George Square, Edinburgh EH8 9XD, UK

TISSUE factor is a cell-surface glycoprotein receptor which initiates the blood coagulation cascade after vessel injury by interacting with blood clotting factor VII/VIIa and which is implicated in various pathological processes¹. When bound to tissue factor, factor VII is readily converted to the active protease factor VIIa by trace amounts of factors Xa, IXa or VIIa. Human tissue factor consists of 263 residues, the first 219 of which comprise the extracellular region². We have determined the crystal structure of the extracellular region at a resolution of 2.2 Å. Tissue factor consists of two immunoglobulin-like domains associated through an extensive, novel, interdomain interface region. The binding site for factor VII lies at the interface region and involves residues from domain 1 and an extended loop (binding 'finger') of domain 2. This is the first reported structure of a representative of the class 2 cytokine receptor family, which also includes interferon-α, interferon-γ (refs 2, 3) and interleukin-10 (ref. 4) receptors.

Residues 1–219 of tissue factor (TF₂₁₉) were expressed in *Escherichia coli*, purified and crystallized as described previously⁵. TF₂₁₉ was fully functional in assays for soluble tissue-factor activity⁶ and bound to factor VII in ligand blots⁷. The structure of TF₂₁₉ was solved by multiple isomorphous

Received 4 May; accepted 29 June 1994.

1. Noller, H. F. et al. in *The Ribosome: Structure, Function, and Evolution* (ed. Hill, W. E. et al.) 73–92 (Am. Soc. Microbiol., Washington DC, 1990).
2. Noller, H. F. in *The RNA World* (eds Gesteland, R. F. & Atkins, J. F.) 137–184 (Cold Spring Harbor Laboratory Press, New York, 1993).
3. Stern, S., Powers, T., Changchien, L.-M. & Noller, H. F. *Science* **244**, 783–790 (1989).
4. von Ahsen, U., Davies, J. & Schroeder, R. *Nature* **353**, 368–370 (1991).
5. von Ahsen, U., Davies, J. & Schroeder, R. *J. molec. Biol.* **226**, 935–941 (1992).

§ Present addresses: Haemostasis Research Group, MRC Clinical Sciences Centre, Royal Postgraduate Medical School, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK (D.M.A.M., D.P.O.B., E.G.D.T.); Principal's Office, University of Sterling, Sterling FK9 4LA, UK (A.M.).

|| To whom correspondence should be addressed.

TABLE 1 Statistics of crystallographic structure determination

Diffraction data and MIR phasing		Resolution $d_{\min}(\text{\AA})$	No. of measurements	No. of unique reflections (completeness)	R_{merge} (I)	MFID (F)	Phasing power
Type A crystals 44.76 × 44.76 × 231.18 Å							
Native 1	(IP-Lab)	3.0	24,308	4,969 (94%)	12.4%	—	—
Trimethyllead	(IP-SRS)	2.95	44,077	5,329 (96%)	10.4%	22.1%	0.8
Cis-Pt	(IP-Lab)	3.5	22,443	3,351 (98%)	10.1%	20.4%	1.0
Type B crystals 45.40 × 45.40 × 230.80 Å							
Native 2	(IP-SRS)	3.1	36,384	4,881 (99%)	8.1%	—	—
Native 3	(IP-KEK)	2.2	42,892	12,292 (93%)	5.4%	—	—
Native 4	(IP-KEK)	2.3	38,704	10,802 (93%)	5.5%	—	—
Native 2 + 3 + 4		2.2		12,666 (96%)	6.5%	—	—
Baker's	(IP-Lab)	3.9	12,003	2,397 (93%)	10.1%	11.9%	0.9
K ₂ HgI ₄	(IP-Lab)	3.5	10,711	3,022 (87%)	11.6%	27.0%	0.5
UO ₂ (NO ₃) ₂	(IP-SRS)	3.0	34,813	4,731 (87%)	10.2%	19.0%	1.1

TF₂₁₉, previously known as TF₂₂₀, was expressed, purified and crystallized as described⁵. The crystals belong to the tetragonal space group P₄₁2₁2, with one molecule per asymmetric unit. Diffraction datasets were collected in-house, at the SRS-Synchrotron in Daresbury, UK (station 9.5), and at the Photon-Factory (KEK) in Japan. In-house and SRS data were collected using a Mar-Research imaging-plate system with a plate diameter of 18 cm. The trimethyllead dataset was collected at $\lambda = 0.947$ Å to optimize the anomalous scattering signal for lead. Data at the Photon-Factory were collected using the Weissenberg method on station BL6A2 (ref. 22) at $\lambda = 0.97$ Å using BASIII imaging plates scanned off-line with Fuji BA100 IP scanners. Data were processed and scaled with the programs DENZO and SCALEPACK²³. ($R_{\text{merge}} = \sum |I_{\text{obs}} - I_{\text{avg}}| / \sum I_{\text{avg}}$; MFID, mean fractional isomorphous difference; Phasing power, mean value of the heavy-atom structure factor amplitudes divided by the residual lack-of-closure error, calculated with MLPHARE²⁴.) As tissue factor crystals varied in unit cell dimensions and diffraction patterns, individual datasets were grouped as types A or B. The MFID between native 1 and 2 was 26.35%. Five useful heavy-metal derivatives were prepared by soaking experiments: 1, trimethyllead acetate²⁵, 30 mM for 8 days; 2, *cis*-dichlorodiamineplatinum (II) (Cis-Pt), 10 mM (saturated) for 2 days; 3, dithiothreitol, 10 mM for 20 h, followed by 1,4-diacetoxymercuro-2,3-dimethoxybutane (Baker's), 2 mM (saturated) for 16 h; 4, K₂HgI₄, 2 mM in 10 mM KI for 19 h; and 5, UO₂(NO₃)₂, 10 mM for 3 days. Heavy-metal binding sites were determined by Patterson and difference Fourier methods. Phases were calculated with PHASE and REFINER (P. Shaw, unpublished programs) as well as MLPHARE²⁴. Electron density maps for types A and B crystals were calculated to 4.0 Å and then phase-extended to 3.0 Å resolution incorporating MIR data and improving the phases by solvent flattening with the program GAP (D.I.S. and J. Grimes, unpublished data). These two independent maps were superimposed and averaged with GAP and an initial C_α model was fitted into density using the program FRODO²⁶. From this model a poly-Ala chain was constructed with the program CALPHA (R. Esnouf, unpublished program). The binding site for Baker's was identified as the Cys 186–Cys 209 disulphide bridge and from this position the chain could be traced without ambiguity. Refinement used simulated annealing with the program XPLOR²⁷. Several rounds of model rebuilding and refinement led to the present model, which includes a bulk solvent correction²⁷, residues 3–213 (2,076 non-hydrogen atoms) and 40 ordered water molecules, has a crystallographic *R*-factor of 21.8% for all data from 15.0 to 2.2 Å (12,623 reflections; $R = 21.2\%$ for $F > 3\sigma$, 11,585 reflections). The r.m.s. deviation from ideal values is 0.015 Å for bond lengths and 2.01° for bond angles. The mean *B*-factor for main-chain atoms is 47.7 Å² (for all atoms 48.9 Å²) and the r.m.s. *B* deviation for bonded main-chain atoms 3.6 Å². Atomic coordinates will be deposited in the Brookhaven Protein Data Bank.

replacement as described in Table 1 legend. The current model has an *R*-factor of 21.8% on all data to 2.2 Å resolution, with good stereochemistry (see Table 1). TF₂₁₉ consists of two immunoglobulin-like domains connected by a single polypeptide linker (Pro 102 to Asn 107). Each domain is formed by two antiparallel β -sheets with immunoglobulin superfamily (IgSF)-type C2 topology⁸ (Figs 1 and 2), as found in other cell-surface receptors; for example, in both domains of the human growth hormone receptor⁹, in domain 2 of the cell adhesion molecule CD2 (ref. 10) and in domains 2 (refs 11, 12) and 4 (ref. 13) of CD4. The root-mean-squared (r.m.s.) deviations between human growth hormone receptor and TF₂₁₉ of structurally equivalent C_α positions are 1.8 Å for 61 residues of domain 1 and 1.4 Å for 80 residues of domain 2. The difference in inter-domain hinge angle is 45°, with TF₂₁₉ adopting a more 'upright' configuration, as observed in CD2 and CD4.

Tissue factor has two disulphide bridges at positions 49–57 and 186–209 (ref. 14) (Fig. 2). The first disulphide bridge links strand C' with strand E in the same way as the corresponding disulphide in the human growth hormone receptor⁹. The other disulphide bridge, which is required for function¹⁵ and is characteristic of the cytokine class 2 receptor family, lies at the very end of domain 2 and links adjacent strands F and G. The sulphur atoms of this S–S bond point away from the surface of the β -sheet towards the solvent. (In our crystals this bond could be cleaved with dithiothreitol as described in Table 1 legend.) This is the first example of such an S–S bond positioned at the surface rather than internal to the β -sandwich of the IgSF domain.

Beyond residue 210 the structure is less well ordered and we cannot find interpretable electron density for the last six residues (214–219). In full-length tissue factor, these residues form the connection between the extracellular region and the trans-membrane domain, the so-called stalk region⁸. Flexibility of the stalk region would allow the whole extracellular domain to adopt various orientations with respect to the membrane surface while the Cys 186–Cys 209 disulphide bridge would ensure the structural stability of this part of domain 2.

The most interesting aspect of the structure is the extent and nature of the interface region between the two IgSF domains. The connections between several strands contribute to this region through either short helices or extended loops. We observe two short α -helices, α_1 and α_2 , a helical turn at Tyr 103–Glu 105 and a 3_{10} -helical turn at Lys 149–Asp 150. Two extended loops protrude from the interface region at either side of the molecule: residues Ile 193 to Arg 200 connecting strand F to strand G (domain 2) and, most prominently, residues Thr 132 to Ser 142, which connect strand B to strand C (domain 2) and project like a finger away from the protein surface. The domain interface has a large hydrophobic core comprised of 11 residues (Phe 19, Val 64, Val 67, Tyr 103, Leu 108, Leu 133, Val 134, Leu 143, Val 146, Phe 147 and Val 198). The overall surface area buried at the domain interface is approximately 1,000 Å², slightly greater than that buried between domains 1 and 2 of CD4. Given the extent of this region, it seems likely that in tissue factor the connection between domains 1 and 2 is rigid, as it is in CD4 (refs 11, 12).

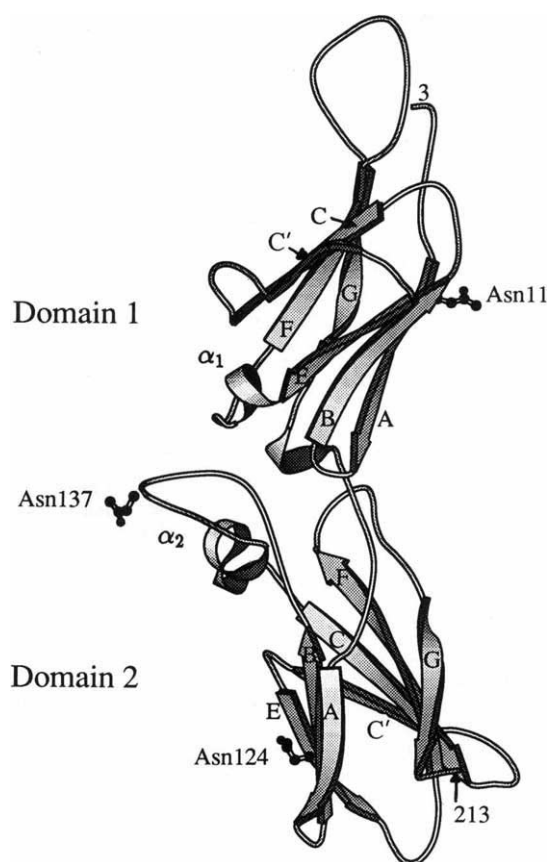


FIG. 1 MOLSCRIPT²⁸ representation of TF₂₁₉. Strands of domain 1 and 2 are labelled A to G; also marked are the two α -helices and three asparagine side chains that provide the glycosylation sites. Numbers 3 and 213 indicate the first and last ordered residue of the structure. The orientation of the molecule is such that the phospholipid membrane would be at the bottom.

Tissue factor has been the subject of extensive mutagenesis experiments¹⁶⁻²⁰. Mapping residues identified as important in the binding of tissue factor to factor VII¹⁶ to our structure highlights a distinct contiguous region in the cleft formed between the finger and strands B and E of domain 1 (Fig. 3). All these residues are well ordered in our structure. Arg 135 and Phe 140

are located on the finger and their side chains point towards domain 1. In the crystal, this binding finger is involved in crystal contacts to another molecule and is only partly exposed to solvent. In solution, the binding finger would be completely exposed to solvent, an unusual environment for a Phe residue. Interestingly, in CD4 the residues that are important for binding to the envelope protein gp120 of the human immunodeficiency virus (HIV) also include an exposed phenylalanine as well as an arginine residue^{11,12}. It seems that only one side of the tissue-factor finger is involved in factor VII binding, as residues such as Arg 136, which point in the opposite direction, were found to have no influence on binding¹⁶. In the human growth hormone/receptor complex⁹ the BC loop of domain 2 of the receptor mediates the growth hormone binding; however, the other structural elements that contribute to the binding region seem to be rather different from those in tissue factor.

There are three *N*-linked glycosylation sites in human tissue factor (Fig. 1) and the factor VII binding site is bordered by two of these: Asn 11 on strand A of domain 1 and Asn 137 at the tip of the binding finger. In our structure, these side chains point away from the binding site. Thus, any carbohydrate attached to these residues is unlikely to mediate or interfere with the binding. This is consistent with data which indicate that factor VII binding is unaffected by the glycosylation state of tissue factor²¹.

We note a cluster of hydrophobic residues on the surface of the C'CFG β -sheet of domain 1 involving residues Val 33, Phe 50, Tyr 51, Phe 76, Tyr 78, Pro 92 and Tyr 94. Some of these residues are involved in hydrophobic interactions with a symmetry-related molecule (Phe 50 to Pro 79, Tyr 51 to Tyr 51 and Tyr 51 to Tyr 78). In the crystal the long FG loop of one molecule is packed against strand C' of another molecule. In solution, this hydrophobic patch is exposed to solvent, as two of the three chymotryptic cleavage sites lie within this region⁷. The function of this hydrophobic cluster is unclear.

Mutagenesis experiments¹⁷⁻²⁰ have identified tissue-factor residues 154-174 as being important in the activation of factor X by the tissue-factor/factor VIIa complex. These residues cluster near the base of domain 2 on strands C, C' and E (Fig. 3a). This activation site is therefore positioned on the opposite face of the molecule to the factor VII binding site at a position which *in vivo* would be close to the phospholipid cell surface. Within this region, Lys 165 and Lys 166 are essential for cofactor function^{17,18}. These residues lie at the very bottom of domain 2 and form the beginning of strand C'. They are partly exposed to solvent and point in opposite directions. However, the role of the phospholipid membrane and the precise nature of the

FIG. 2 Amino-acid sequence of TF₂₁₉ with the positions of β -strands and α -helices indicated and disulphide bridges marked.

```

1          < A > < B > < C > 40
SGTNTVAAY  NLTWKSTNFK  TILEWEPKPV  NQVYTVQIST

41          < C' > < E > <  $\alpha$ 1 > < F > 80
KSGDWKSKCF  YTTDTECDLT  DEIVKDVKQT  YLARVFSYPA

81          < G > <  $\alpha$  $\alpha$  > < A > 120
GNVESTGSAG  EPLYENSPEF  TPYLETNLGG  PTIQSFEQVG

121         < B > <  $\alpha$ 2 > < 3 $\alpha$  > < C > 160
TKVNVTVDEE  RTLVRNNTF  LSLRDVFGKD  LIYTLYYWKS

161         < C' > < E > < F > < 200
SSSGKKTAKT  NTNEFLIDVD  KGENYCFSVQ  AVIPSRTVNR

201         G >
KSTDSPVECM  GQEKGEFRE

```

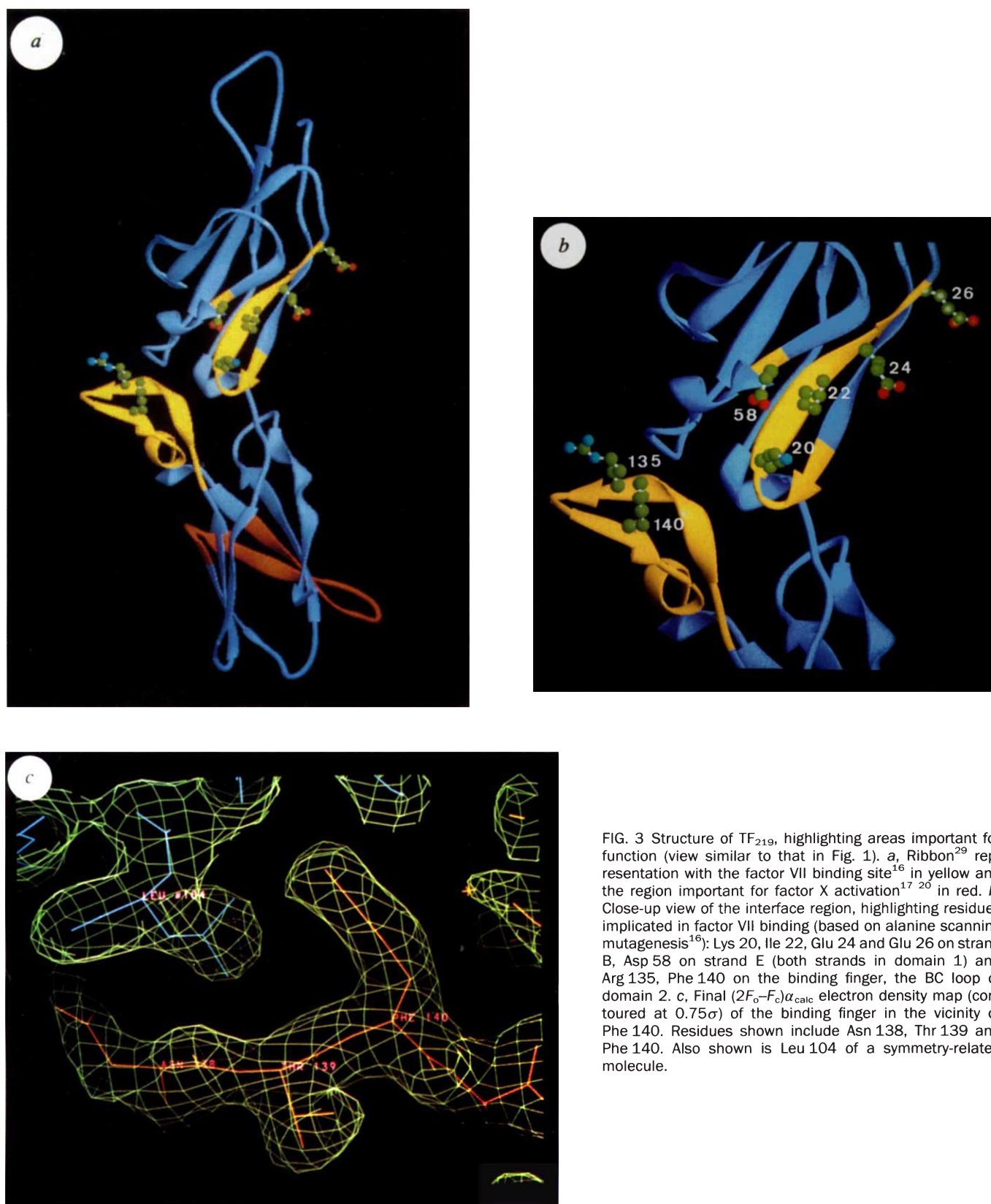


FIG. 3 Structure of TF₂₁₉, highlighting areas important for function (view similar to that in Fig. 1). *a*, Ribbon²⁹ representation with the factor VII binding site¹⁶ in yellow and the region important for factor X activation^{17, 20} in red. *b*, Close-up view of the interface region, highlighting residues implicated in factor VII binding (based on alanine scanning mutagenesis¹⁶): Lys 20, Ile 22, Glu 24 and Glu 26 on strand B, Asp 58 on strand E (both strands in domain 1) and Arg 135, Phe 140 on the binding finger, the BC loop of domain 2. *c*, Final $(2F_o - F_c)\alpha_{calc}$ electron density map (contoured at 0.75σ) of the binding finger in the vicinity of Phe 140. Residues shown include Asn 138, Thr 139 and Phe 140. Also shown is Leu 104 of a symmetry-related molecule.

interactions between factor X and the tissue-factor/factor VIIa complex are as yet unknown. □

Received 3 June; accepted 20 July 1994.

1. Ruf, W. & Edgington, T. S. *FASEB J.* **8**, 385–390 (1994).
2. Edgington, T. S., Mackman, N., Brand, K. & Ruf, W. *Thromb Haemostasis* **66**, 67–79 (1991).
3. Bazan, J. F. *Proc. natn. Acad. Sci. U.S.A.* **87**, 6934–6938 (1991).
4. Ho, A. S. Y. et al. *Proc. natn. Acad. Sci. U.S.A.* **90**, 11267–11271 (1993).

5. Boys, C. W. G. et al. *J. molec. Biol.* **234**, 1263–1265 (1993).
6. Martin, D. M. A., O'Brien, D. P., Tuddenham, E. G. D. & Byfield, P. G. H. *Biochemistry* **32**, 13949–13955 (1993).
7. O'Brien, D. P., Anderson, T. S., Martin, D. M. A., Byfield, P. G. H. & Tuddenham, E. G. D. *Biochem. J.* **292**, 7–12 (1993).
8. Jones, E. Y. *Curr. Opin. struct. Biol.* **3**, 846–852 (1993).
9. de Vos, A. M., Ultsch, M. & Kossiakoff, A. A. *Science* **255**, 306–312 (1992).
10. Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. & Stuart, D. I. *Nature* **360**, 232–239 (1992).
11. Wang, J. et al. *Nature* **348**, 411–418 (1990).
12. Ryu, S. E. et al. *Nature* **348**, 419–426 (1990).

12. Ryu, S. E. et al. *Nature* **348**, 419–426 (1990).
13. Brady, R. L. et al. *Science* **260**, 979–983 (1993).
14. Bach, R., Königsberg, W. H. & Nemerson, Y. *Biochemistry* **27**, 4227–4231 (1988).
15. Rehemtulla, A., Ruf, W. & Edgington, T. S. *J. biol. Chem.* **266**, 10294–10299 (1991).
16. Ruf, W., Schullek, J. R., Stone, M. J. & Edgington, T. S. *Biochemistry* **33**, 1565–1572 (1994).
17. Roy, S., Hass, P. E., Bourell, J. H., Henzel, W. J. & Vehar, G. A. *J. biol. Chem.* **266**, 22063–22066 (1991).
18. Ruf, W., Miles, D. J., Rehemtulla, A. & Edgington, T. S. *J. biol. Chem.* **267**, 6375–6381 (1992).
19. Ruf, W., Miles, D. J., Rehemtulla, A. & Edgington, T. S. *J. biol. Chem.* **267**, 22206–22210 (1992).
20. Gibbs, C. S., McCurdy, S. N., Leung, L. L. K. & Paborsky, L. R. *Blood, Suppl. 1*, **82**, 451a, Abstr. 1788 (1993).
21. Paborsky, L. R. et al. *Biochemistry* **28**, 8072–8077 (1989).
22. Sakabe, N. *Nucl. Instrum. Meth. A* **303**, 448–463 (1991).
23. Otwinowski, Z. in *Data Collection and Processing* (eds Sawyer, L., Isaacs, N. & Bailey, S.) 56–62 (SERC Daresbury Laboratory, Warrington, 1993).
24. Otwinowski, Z. in *Isomorphous Replacement and Anomalous Scattering* (eds Wolf, W., Evans, P. R. & Leslie, A. G. W.) 80–85 (SERC Daresbury Laboratory, Warrington, 1991).
25. Holden, H. M. & Rayment, I. *Archs Biochem. Biophys.* **291**, 187–194 (1991).
26. Jones, A. *Meth. Enzym.* **115**, 157–171 (1985).
27. Brünger, A. T. *X-PLOR, Version 3.1, A System for X-ray Crystallography and NMR* (Yale Univ. Press, New Haven, 1992).
28. Kraulis, P. J. *J. appl. Crystallogr.* **24**, 946–950 (1991).

29. Carson, M. J. *appl. Crystallogr.* **24**, 958–961 (1991).

ACKNOWLEDGEMENTS. We thank S. McSweeney (SRS-Daresbury), A. Nakagawa and N. Sakabe (Photon Factory) and R. Esnouf, J.-S. Ren and D. Stammers for help with data collection; D. Higgins and D. Eaton (Genentech) for providing the TF₂₁₉ cell paste; D. Alexeev for help with software; also V. Fülöp, A. N. Barclay, S. J. Davis, R. Robinson, D. L. Bodian, I. D. Campbell and S. Lee. The Oxford Centre for Molecular Sciences is supported by the SERC and MRC; D.M.A.M., D.P.O.B. and E.G.D.T. by MRC; E.Y.J. by a Royal Society University Research Fellowship; and

Structural similarity between the p17 matrix protein of HIV-1 and interferon- γ

Stephen Matthews^{*†}, Paul Barlow^{*}, Jonathan Boyd[†], Geoff Barton[‡], Robert Russell[‡], Helen Mills[§], Mark Cunningham[§], Nicola Meyers[§], Nigel Burns[§], Nigel Clark^{*}, Susan Kingsman^{*}, Alan Kingsman^{*} & Iain Campbell^{*†||}

^{*} Department of Biochemistry, [†] Oxford Centre for Molecular Sciences and [‡] Laboratory of Molecular Biophysics, University of Oxford, Oxford OX1 3QU, UK

[§] British Biotechnology Ltd, Watlington Road, Oxford OX4 5LY, UK

THE human immunodeficiency virus (HIV) matrix protein, p17, forms the outer shell of the core of the virus, lining the inner surface of the viral membrane^{1–4}. The protein has several key functions. It orchestrates viral assembly via targeting signals that direct the gag precursor polypeptide, p55, to the host cell membrane^{1,5–7} and it interacts with the transmembrane protein, gp41, to retain the env-encoded proteins in the virus⁸. In addition, p17 contains a nuclear localization signal that directs the preintegration complex to the nucleus of infected cells⁹. This permits the virus to infect productively non-dividing cells, a distinguishing feature of HIV and other lentiviruses. We have determined the solution structure of p17 by nuclear magnetic resonance (NMR) with a root-mean square deviation for the backbone of the well-defined regions of 0.9 Å. It consists of four helices connected by short loops and an irregular, mixed β -sheet which provides a positively charged surface for interaction with the inner layer of the membrane. The helical topology is unusual; the Brookhaven protein database contains only one similar structure, that of the immune modulator interferon- γ .

To produce sufficient protein for NMR analysis, p17 was expressed in *Escherichia coli* as a glutathione *S*-transferase fusion. The p17 coding sequence came from the HXB2 proviral clone. Following glutathione affinity purification, thrombin cleavage and ion-exchange chromatography, 20–30 mg of pure

(>95%) protein was recovered from a 5-l fermentation. Uniformly ¹⁵N-labelled and ¹⁵N-leucine-labelled p17 gave sufficiently good spectra to allow a full backbone ¹H assignment using ¹⁵N-¹H heteronuclear multiple quantum coherence nuclear Overhauser (HMQC-NOESY) and Hartmann-Hahn (HMQC-HOHAHA) spectroscopy^{10,11}. The side-chain assignments were completed using ¹³C-¹H HMQC-NOESY and ¹H-¹³C-¹³C-¹H-total correlated spectroscopy¹² (HCCH-TOCSY) on a uniformly labelled ¹³C sample.

The structures were calculated on the basis of 1,067 NOE distance restraints, 75 ϕ dihedral angle restraints and 26 hydrogen-bond distance restraints. The calculations were made within the programme XPLOR using a dynamical simulated annealing protocol^{13,14}. Fifty structures were calculated, of which 26 contained no distance violations greater than 0.5 Å. The family of 26 structures is shown in Fig. 1a superimposed on the mean coordinate position for the backbone atoms of residues 15–110. For these residues, the average root-mean-square deviation (r.m.s.d.) from the average structure is 1.34(±0.3) Å for the backbone atoms and 1.72(±0.4) Å for all atoms; for the regions with defined secondary structure these values are 0.92(±0.2) Å and 1.44(±0.3) Å, respectively.

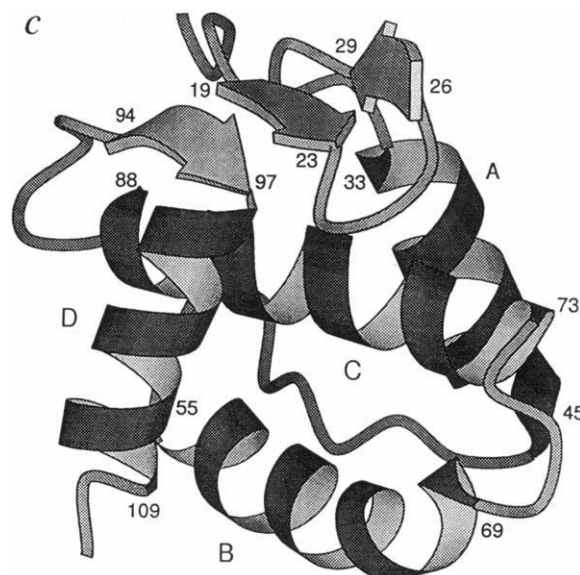
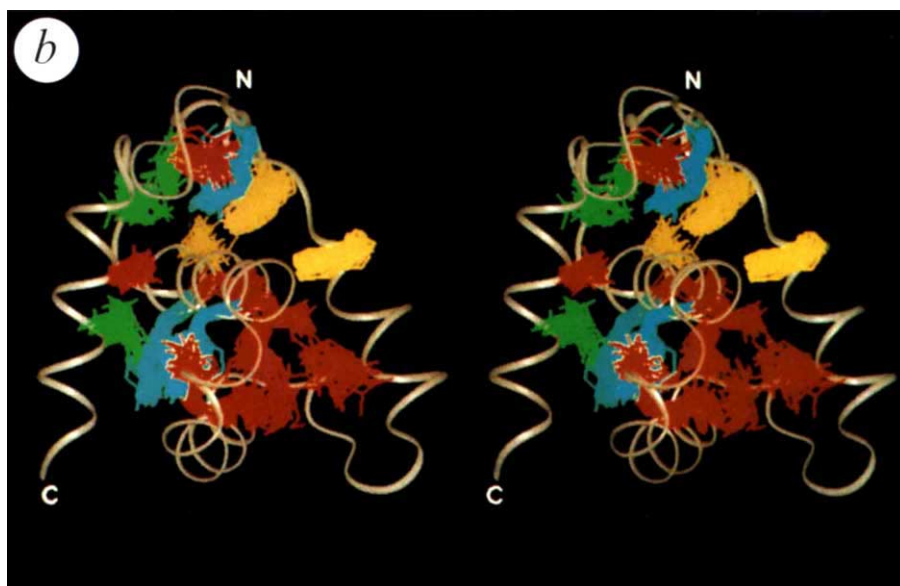
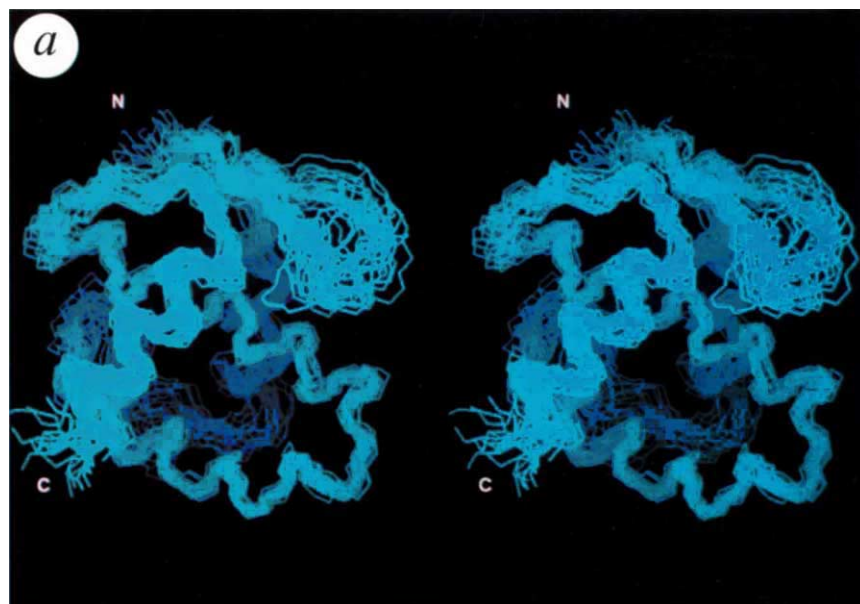
The structure has a compact fold which contains four helices joined by short loops and a triple-stranded, irregular, mixed β -sheet. Helices B and C form an antiparallel coiled coil, which lies in the centre of the molecule. Helices A and D lie approximately parallel to each other on either side of the coiled coil. This arrangement gives rise to close packing of adjacent helices where the interhelical angles are 82°, 168° and 78° for AB, BC and CD, respectively. All the helices are accessible to the solvent and highly amphipathic, with the exception of helix C which runs through the centre of the hydrophobic core. Figure 1b, shows the radial distribution of the hydrophobic residues around this buried helix. Two regions (residues 19–23 and 26–29) towards the amino terminus, together with the residues between helices C and D (94–97), form three strands of an irregular, mixed β -sheet which constitutes a platform on 'top' of the structure. The lack of ¹H-¹H NOE data and a strong negative ¹⁵N-¹H heteronuclear NOE¹⁵ experienced by the N-amide resonances of the carboxy-terminal residues (data not shown) strongly suggest that the C-terminal 20 amino acids do not adopt a rigid conformation in solution. There is evidence for a turn centred about G10 and G11 (single-letter amino-acid code) within the N-terminal 14 amino acids, but it is not well defined with respect to the rest of the molecule.

An automated search of the Brookhaven protein structure database revealed similarity between p17 and the topology of the interferon- γ monomer^{16–18} (Fig. 2). As Fig. 2b shows, there is essentially no sequence similarity between p17 and interferon- γ . Forty-one Ca atoms from the helical regions in p17 (A, B, C and D) superimpose on the equivalent regions (helices A, C, D and E) in interferon- γ with a r.m.s.d. of 2.6 Å. There are, however, two significant differences between the two structures. First, the β -sheet structure seen in p17 is not present in interferon- γ . Second, a short helical region exists within the long loop joining the major helices A and C of interferon- γ ; in p17 this loop is significantly shorter and lacks any recognizable secondary structure. Interferon- γ is functional as a homodimer, in which there is unique interdigitizing of the helices across the subunit faces. The C termini of the monomers form helical arms which fit into a cleft between helices A and C of the adjacent monomer; they are not well structured^{19,20}. It is conceivable that p17 dimerizes in an analogous way during viral assembly, as it also has a similar helical topology and contains a flexible C terminus. Extended arm- and helix-mediated interactions are common features in viral architecture²¹. The sequences of the simian immunodeficiency virus (SIV), HIV-1 and HIV-2 matrix proteins are highly homologous with each other. This strongly implies that any conclusions based on the structure of the HIV-1 matrix protein also hold true for all strains of HIV and SIV.

|| To whom correspondence should be addressed.

FIG. 1 *a*, Stereoview of backbone traces for the final 26 structures of HIV-1 p17 (residues 15–110) superimposed on the average coordinate position. *b*, Stereoview of the superposition of the principal hydrophobic side chains of the 26 structures. Leucine sidechains are shown in red, tyrosine in blue, isoleucine in green, and both valine and tryptophan are in yellow. The backbone of the average structure is represented by a white ribbon. *c*, Schematic representation of HIV-1 p17 showing the relative positions of the α -helices and β -sheet. The diagram was produced using the programme MOLSCRIPT²⁴.

METHODS. HIV-1 p17 was expressed using a pGEX-2T plasmid in *E. coli* JM103 cells. Labelled samples were obtained from growth of the *E. coli* on minimal media containing either $(^{15}\text{NH}_4)_2\text{SO}_4$ as the sole nitrogen source or $\text{U-}^{13}\text{C}$ -glucose as the sole carbon source. The defined media for specifically labelling the leucine residues with ^{15}N have been described elsewhere²⁵. All NMR spectra were recorded at 35 °C on 2 mM protein samples which contained 30 mM phosphate buffer at pH 6.0 and 15 mM NaCl. Three spectrometers were used, operating at proton frequencies of 500, 600 and 750 MHz. The structures were calculated on the basis of 1,067 NOE distance constraints, 75 ϕ dihedral angle restraints and 26 hydrogen-bond distance restraints. The NOE restraints were composed of 219 intra-residue, 304 sequential (residue *i* to residues *i*+1), 280 medium range (residue *i* to residues $1 < i \leq 4$) and 264 long-range (residue *i* to residues $i > 4$) connectivities. The distance restraints were categorized into three groups on the basis of estimated NOE cross-peak intensity: strong, 2.7 Å; medium, 3.8 Å; weak, 5.0 Å. Standard values for pseudo atom corrections²⁶ were added to the NOE distance constraints where appropriate. The ϕ restraints were derived from $^3J(\text{NH}_\alpha)$ by line-fitting traces from two-dimensional ^{15}N - ^1H HMQC-*J* experiments²⁷. Hydrogen-bonded NH groups were identified in helical regions by the presence of NH resonances 6 h after dissolving in D_2O . The final values for each term in the experimental terms in the force field used during simulated annealing were $F_{\text{NOE}} = 31.5 \pm 9 \text{ kcal mol}^{-1}$, $F_{\text{Cdihed}} = 2.9 \pm 1 \text{ kcal mol}^{-1}$, with force constants of 50 and 200 $\text{kcal mol}^{-1} \text{ rad}^{-2}$, respectively. The average Lennard-Jones energy for the selected structures is $-385.5 \pm 20 \text{ kcal mol}^{-1}$, which is consistent with good non-bonded contacts. The r.m.s.d. from the experimental restraints and idealized geometry are as follows: NOE distances $0.023 \pm 0.003 \text{ Å}$, dihedral restraints $0.62 \pm 0.1^\circ$, bonds $0.0027 \pm 0.0002 \text{ Å}$, angles $0.49 \pm 0.03^\circ$ and improper angles $0.36 \pm 0.03^\circ$. The coordinates will be deposited in the Brookhaven protein database.



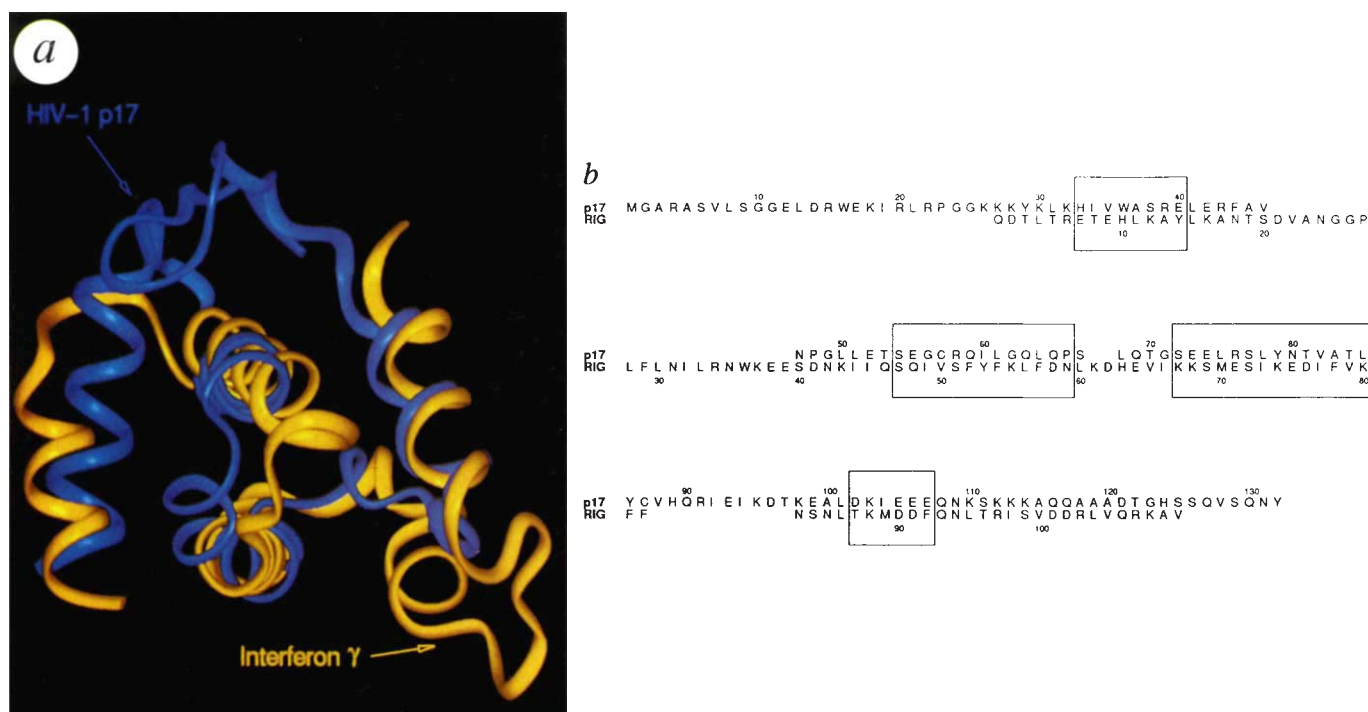


FIG. 2 **a**, A ribbon representation of the superimposition of the Ca atoms judged to be structurally equivalent within p17 and rabbit interferon- γ ¹⁶. For clarity the C termini are not shown. **b**, The sequence alignment for p17 and rabbit interferon- γ . The 41 equivalent positions are indicated by boxes. These do not, of course, correspond exactly to the limits of the secondary structure regions shown in Fig. 1c. METHODS. A database of 391 unique protein three-dimensional structural domains was searched by overlaying the sequences of p17 with

every 5th residue in each database structure. Residues aligned by this procedure were used to derive an initial fit of p17 and the database structure. The initial fit was then refined using the protein-structure comparison algorithm described in ref. 28. The topology of p17 was judged using this method to be closest to that of the fold of interferon- γ . 41 equivalent Ca atoms were used to derive the superimposition shown, and gave a r.m.s.d. of 2.6 Å.

The irregular β -sheet is a prominent feature of p17. A glycine residue at the N terminus of p17 is myristoylated in the virus and many studies have established that this modification is essential for membrane targeting and virus assembly. It seems likely that this modified residue will be buried in the viral membrane. The solvent-exposed side of the β -sheet provides a surface which could associate with the inner face of the membrane, with several basic side chains (K18, R20, R22, K26–K28, K30, K32, K95) available for interaction with phospholipids headgroups. This is supported by mutagenesis studies in which mutations that alter the charge distribution have dramatic effects on virus assembly^{6,7} and is consistent with the notion that the myristoylated glycine and the basic patch from residues 18 to 32 form a bipartite

autonomous membrane-targeting signal²². This type of N-terminal basic region is common among the C-type retroviral matrix proteins. In HIV, the basic patch is also a major epitope for cytotoxic T lymphocytes²³ and a nuclear localization signal for the preintegration complex⁹. In addition, it is possible that the β -sheet interacts with gp41 to retain the viral envelope proteins within the particle⁸.

We believe our study represents the first report of the structure of a retroviral matrix protein. The structural similarity of the HIV matrix to the immune modulator, interferon- γ , is intriguing and may provide a useful model for p17 dimerization as an initial step in virus assembly. It remains to be seen whether the p17/interferon- γ relationship is significant for the induction of immunodeficiency in HIV-infected individuals. □

Received 18 May; accepted 27 July 1994.

1. Spearman, P., Wang, J., Heyden, N. V. & Ratner, L. *J. Virol.* **68**, 3233–3242 (1994).
2. Gelderblom, H. R., Özel, M. & Pauli, G. *Archs Virol.* **106**, 1–13 (1989).
3. Nermut, M. V. et al. *Virology* **198**, 288–296 (1994).
4. Arnold, E. & Arnold, G. F. *Adv. Virus Res.* **39**, 1–58 (1991).
5. Bryant, M. & Ratner, L. *Proc. natn. Acad. Sci. U.S.A.* **87**, 523–527 (1989).
6. Yuan, X., Yu, X., Lee, T. & Essex, M. J. *J. Virol.* **67**, 6387–6394 (1993).
7. González, S. A., Affranchino, J. L., Gelderblom, H. R. & Burny, A. *Virology* **194**, 548–556 (1993).
8. Yu, X., Yuan, X., Matsuda, Z., Lee, T. & Essex, M. J. *J. Virol.* **66**, 4966–4971 (1992).
9. Bukrinsky, M. I. et al. *Nature* **365**, 666–669 (1993).
10. Marion, D. et al. *Biochemistry* **28**, 6150–6156 (1989).
11. Driscoll, P. C., Clore, G. M., Marion, D., Wingfield, P. T. & Gronenborn, A. M. *Biochemistry* **29**, 3542–3556 (1990).
12. Bax, A., Clore, G. M. & Gronenborn, A. M. *J. magn. Reson.* **88**, 425–431 (1990).
13. Nilges, M., Gronenborn, A. M., Brünger, A. T. & Clore, G. M. *Protein Engng* **2**, 27–38 (1988).
14. Brünger, A. T. *XPLOR Manual Ver 3.1* (Yale Univ. Press, New Haven, 1993).

15. Main, A. L., Harvey, T. S., Baron, M., Boyd, J. & Campbell, I. D. *Cell* **71**, 671–678 (1992).
16. Samudzi, C. T., Buston, L. E. & Rubin, J. R. *J. biol. Chem.* **266**, 21791–21797 (1991).
17. Ealick, S. E. et al. *Science* **252**, 698–702 (1991).
18. Farrer, M. A. & Schreiber, R. D. *A. Rev. Immun.* **11**, 571–611 (1993).
19. Lundell, D. et al. *Protein Engng* **4**, 335–341 (1991).
20. Grzesiek, S. et al. *Biochemistry* **31**, 8180–8190 (1992).
21. Stuart, D. *Curr. Opin. struct. Biol.* **3**, 167–174 (1993).
22. Zhou, W., Parent, L. J., Wills, J. W. & Resh, M. D. *J. Virol.* **68**, 2556–2569 (1994).
23. Johnson, R. P. et al. *J. Immun.* **147**, 1512–1521 (1991).
24. Kraulis, P. J. *J. appl. Crystallogr.* **24**, 946–950 (1991).
25. Matthews, S. J., Jandu, S. K. & Leatherbarrow, R. J. *Biochemistry* **32**, 657–662 (1993).
26. Wüthrich, K., Billeter, M. & Braun, W. *J. molec. Biol.* **169**, 949–961 (1983).
27. Kay, L. A. & Bax, A. *J. magn. Reson.* **86**, 110–126 (1990).
28. Russell, R. B. & Barton, G. J. *Proteins: Struct. Funct. Genet.* **14**, 309–323 (1992).

ACKNOWLEDGEMENTS. We thank N. Wood, G. Wingfield and S. Cole for help with *E. coli* fermentation, K. Jones and S. Champion for protein purification; and P. Cannon and I. Jones for discussion. This work is supported by the MRC AIDS Directed Programme. This is a contribution from the Oxford Centre for Molecular Sciences supported by the SERC and MRC.

Computer age

Software and hardware options featured this week include programs for data acquisition, management and analysis, imaging and molecular biology, and a selection of new computer systems.

PICK up a copy of SciTech's latest *Software for Science* 96-page catalogue of scientific and technical software for DOS, Windows, Macintosh and UNIX workstations (*Reader Service No. 100*). Three hundred new products are featured — all listed by subject area. If the product you are looking for is not listed, SciTech says it can also tap into its product database of over 3,000 scientific and technical software tools.

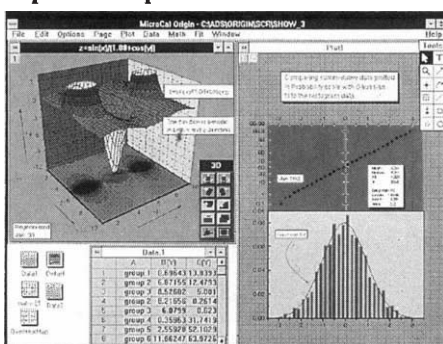
Derwent Direct introduces Patent Scan, a collection of **US patent information on CD-ROM**, which provides users with access to more than 1.7 million patent records (*Reader Service No. 101*). The standard edition of Patent Scan costs US\$995 and consists of twenty years of patent information spanning the years 1974 to 1993. Patent Scan Update, which costs \$1,195, is a monthly updated cumulative index for the current year. It will contain all bibliographic information, as well as abstracts and exemplary claims on a single CD-ROM. Purchasers will receive a new CD-ROM each month that will contain information to date for the current year with the most recent patent information available. Patent Scan Plus, a \$5,000 ten CD-ROM set containing the text of abstracts and claims from 1974 to 1993, completes the line-up. Each \$500 CD-ROM contains two years of data, so users can purchase the most relevant years or the entire set.

■ Presentation

Last month, Jandel Scientific began shipping US\$995 SigmaSuite for Windows, a **suite of graphics, statistics and image measurement applications** designed to work together on the same data (*Reader Service No. 102*). The new bundle of programs is directed at scientific researchers and engineers, and includes the company's SigmaPlot, SigmaStat and SigmaScan/Image image measurement software. Now users can measure images (in SigmaScan/Image), analyse the data (in SigmaStat) and then present

the results in a graphical form (using SigmaPlot). All three share the same 16,000 column by 65,000 row worksheet and data analysis tools, and the same graphics page. SigmaScan/Image can be used for the measurement of monochrome or colour TIFF, TGA, BMP, PCX and GIF images. Statistical analyses included as part of SigmaStat include *t*-tests, analysis of variance, non-parametric statistics, correlation (linear and nonlinear), regression, rates and proportions, normality and equal variance tests. System requirements include a 386 IBM PC or compatible computer, and at least 8 MB of RAM to run all three programs at once.

Newly available — version 3.5 of Origin, Microcal Software's Windows-based software package designed to acquire, manipulate and present data for scientific and



Windows-based graphics program.

engineering applications (*Reader Service No. 103*). Origin's data acquisition system consists of a family of modules which provide data acquisition support for Origin. These modules include LabData for plug-in boards, Lab-GPIB for IEEE 488 instruments and LabCom for RS232 devices. A real-time module is available for real-time strip charting, calibrations, data storage and other data acquisition tasks. These modules are designed to provide functional building blocks for constructing data acquisition systems without writing any code. In addition, there is a three-dimensional contour module, a peak-fitting module and a file utilities module which extends LabTalk's (Origin's scripting language) ability to handle files.

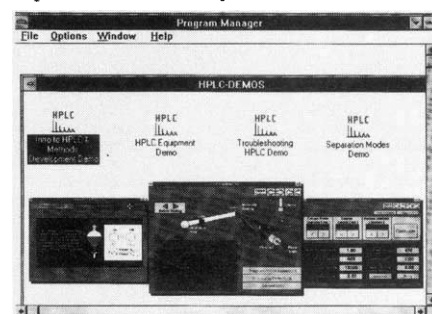
A new version of Kotev Technology's **OPR, optical plot reading software** for converting scanned plots to *x, y* data, is now shipping (*Reader Service No. 104*). OPR offers automatic acquisition of several thousands' points (depending on the resolution of the scanner) from a line (solid or dashed) or scatter graph. Input is provided by any hand-held or flatbed scanner in TIFF format. The program has several digitizing modes:

automatic, semiautomatic and manual. It can process various so-called problematic cases, such as tilted plots and multiple line plots. Heavy grid lines can be removed by two methods. Other useful routines include smoothing and point ordering according to constant *x* intervals. Processed data can be saved as an ASCII file in *.prm format for further processing in any spreadsheet program. System requirements include a 286 IBM PC or compatible computer running DOS. The US\$315 complete version is aimed at researchers in areas such as biotechnology, chemistry, the Earth sciences, physics and also at scientific illustrators. Educational institutions can purchase the package at a reduced cost of \$145.

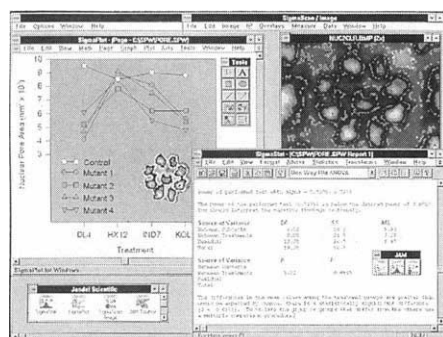
■ Applications

Kalcium is Kinetic's new integrated **package for dynamic intracellular calcium analysis** (*Reader Service No. 105*). Operating under Windows and exploiting 486 or Pentium processors, Kalcium is intended to offer the user a range of image acquisition, processing, analysis and storage options. Images can be acquired from cooled, slow-scan CCD, integrating video or intensified video cameras. Kinetic says that the simultaneous observation of multiple regions of interest enables data to be handled in either image or multipoint photometer mode, with real-time monitoring and display. Storage options include disk arrays, 'fast and wide' SCSI, in addition to interfacing to existing systems. Support is provided for peripheral devices (filter wheels, monochromators, trigger channels and motorized stages).

A new demonstration disk, illustrating five new **HPLC interactive training and troubleshooting software** modules for IBM PCs or compatible computers, is now available from Savant Audiovisuals (*Reader Service No. 106*). Samples from all five programs (developed by Dennis Saunders of the Unocal Science & Technology Division) — *Introduction to HPLC*, *Method Development in HPLC*, *HPLC Equipment* and *Separation Modes of HPLC* — are con-



HPLC troubleshooting/training demo.



SigmaSuite — three programs in one.

PRODUCT REVIEW

tained on the disk. Full versions of the programs feature animated graphics and a chromatographic simulator that shows results of changes in conditions on a synthesized chromatogram when experimental conditions are changed. Experiments are automatically recorded in an electronic notebook, which can be printed.

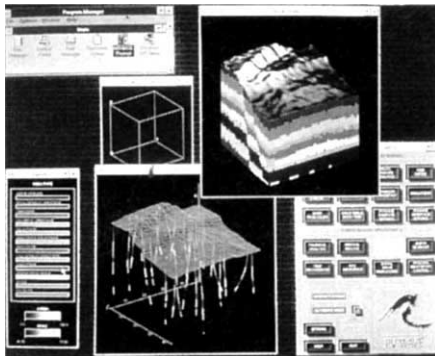
Back in June, Floating Point Systems launched ENVI (Environment for Visualizing Images), an image processing system designed for **multi-spectral analysis of satellite and aircraft remote-sensed data** (Reader Service No. 107). Available on most platforms, including UNIX workstations, IBM or compatible PCs running Microsoft Windows and Apple Macintosh systems, ENVI is designed to allow the user to display and analyse multi-band data sets of any size, data type and spectral depth. It is available as development systems (prices start at UK£8,750 for the UNIX version; £4,475 for the Windows version) or run-time (ENVI-RT) systems. The run-time systems have all ENVI's features but cannot be customized, and are priced at £5,750 for UNIX, £3,350 for Windows and Mac versions. The program (developed by a group of Earth scientists at CSRO Boulder, Colorado) can read a range of data formats, including MSS, TM, SPOT, ERDAS7.5, ERS-1, AVHRR, AVIRIS, GERIS, GEOSCAN and TMS.

What if is a new **molecular modelling software program** for protein engineering, drug design, NMR spectroscopy and crystallography (Reader Service No. 108). The program, which took Gerrit Vriend who is at the European Molecular Biology Laboratory, Heidelberg, Germany seven years to develop, runs on most workstations (Sun, Evans & Sutherland, DEC-Alpha, IBM RS6000, Silicon Graphics) and also 386 IBM PCs or compatible computers. It is designed to allow the user to ask questions that start with "What if..." and then continue, for example, with "... I mutated that valine

into an isoleucine?" The program can be used to help calculate the consequences of such a mutation. To do so, it can use a three-dimensional relational protein database in one, two or three dimensions. It allows for evaluations of mutations in terms of occupied space, van der Waals' contacts, hydrogen bridges, accessible surfaces, etcetera. Over 1,300 options are featured. Its use with a graphics device allows for the interactive manipulation and display of structures.

■ Imaging

Visual Numerics, a developer of **data visualization software** for UNIX-based workstations, has launched a PC version of its flagship product, PV-WAVE (Reader Service No. 109). Named PV-WAVE Personal Edition, the program runs under Windows 3.1 and costs UK£850. PV-WAVE Personal Edition is a 32-bit application for users of Windows who need visually to analyse large amounts of complex data, enabling them to spot trends and anomalies that might otherwise remain hidden. Integrated



PV-WAVE personal edition for the PC.

two- and three-dimensional graphics, surface, vector and other plot types, image and signal processing, flexible data input/output, support for time-series data, and animation are among the functions that assist users in the interpretation of data. Optional modules include PV-WAVE Math (for math-

ematical problem solving) and PV-WAVE Stat (for statistical analyses). Each is priced at £200.

VISTA is a new system from Life Science Resources for automated, **time-lapse digital image acquisition and quantitative image processing** for researchers in the life sciences (Reader Service No. 110). It is



New VISTAs from Life Science Resources.

designed to enable the acquisition of high-resolution digital images from monochrome or colour cameras, at resolutions up to 768 × 576 pixels. The package is available as an 8-bit monochrome/pseudocolour version and a 24-bit true colour version. Supplied as a completely integrated system, VISTA includes hardware and software for image acquisition, a low-light-level camera, graphics interface and display technology, and computer.

Data Translation announces that NIH Image—a Macintosh-based **image processing and analysis application**, developed at the US National Institutes of Health—is now shipping (Reader Service No. 111). The program provides functions like the measurement of objects, intensity analysis, image input and image processing. It is designed to be used with QuickCapture, a framegrabber board for the real-time capture of 8-bit monochrome images and non-real-time capture of 24-bit RGB images on Macintosh computers. The cost for this package is US\$1,295.

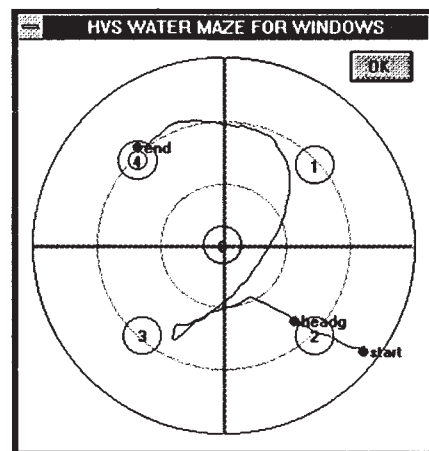
Spyglass has rolled out another addition to its suite of data analysis tools for scientists and engineers in Spyglass Slicer, a **volume metric data analysis tool** for Windows/Windows NT (Reader Service No. 112). Slicer costs US\$695 and enables the user to create three-dimensional colour images. It reads a variety of data and image formats, including HDF (float and byte), netCDF, and binary and ASCII matrices. Additional import flexibility is provided as the program can convert stacks of two-dimensional, HDF, binary and ASCII data files; Slicer can also read HDF, Windows Bitmap, Targa and TIFF image files. After importing the data, Slicer uses ray tracing to render data as volumes, slices or isosurfaces. Full control of alpha values makes it possible to make any object or data value transparent, translucent or opaque. Users can render slices in

ADVERTISEMENTS

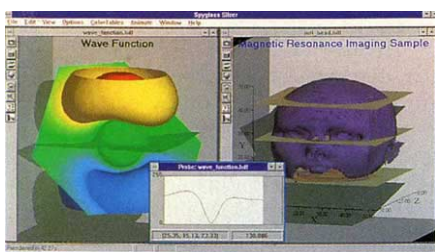
Electrophoresis, DNA, RFLP, TLC, Dots-EUSA, & Image analysis
Universal Software Win/DOS IBM / Mac
 8, 12, 16, & 32-bits TIFF, PCX, ASCII Fax for demo
 To work with:
 1. **Economic Scanners:** HP desktop, Logitech, any video
 2. **Top imagers:** Fuji, Biorad, Mol Dyn, Millipore, & Lasers

--- **Basic4:** 1D, 2D, DNA Seq & Image analysis package
 Automatic mark-read-report. (Mac/Win \$750) DOS \$495
 --- **Main1D:** auto read, map graph, pos, mw, area, % & mg, overlay, Gauss peakfit, +32 functions (Mac/Win \$1,250) \$995
 --- **Advanced1D:** Main1D plus compare 2 or many (ratios, % change, graphs), add, subtract, average, bank, identify, % similarity, (Competition \$18,000) (Mac/Win \$7,500) \$4,995
 --- **Main 2D:** DNA or RFLP (Mac/Win \$1,250) \$995, Archiving \$995
 --- **HP II Scanner:** \$650 **CCD Video-grabber:** \$850 **Logitech:** \$115
 AAB 1020 S. Raymond Ave. #A, Fullerton, CA 92631 USA
 Phn. 714 870-7857 0290 Fax 17148706385, 6384

Reader Service No.5



Reader Service No. 8



Volumetric visualization with Slicer.

orthogonal planes or at arbitrary angles, and isosurfaces can be created for specified values. Additionally, Slicer can create animations from user-specified key frames, and volumes can also be oriented at any arbitrary angle. Variable lighting parameters also give users shading options. Images can be printed to any colour or greyscale Windows-supported printer. They can also be saved as HDF, Windows Bitmap, TIFF, Targa or GIF image file formats.

■ Data management

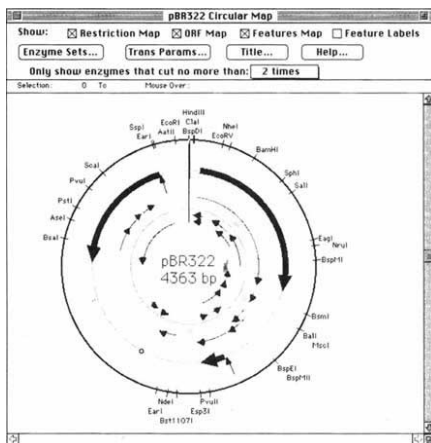
Megalon's new product is ResearchStation, scientific information management software for IBM or compatible personal computers (*Reader Service No. 113*). Based on Windows-standard OLE 2.0 technology, ResearchStation provides an electronic workspace in which scientists can gather, combine and use information in any form: text, images, numerical data, graphics, video and sound. The US\$995 standalone version of the program, which requires a 386/486 PC with Windows 3.1 or higher, 4 MB RAM and 10 MB free hard disk space, is expected to ship within the month; networked versions are expected to be available by November. Developed by Helix Systems, Inc. of Palo Alto, California, ResearchStation is designed to serve both as a personal productivity and research tool for individuals, and as a communication vehicle for collaborative work. The ResearchStation Minder can be used to automate routine and repetitive tasks, and enables users to locate and exchange information contained in internal files, local and remote databases, and other sources.

Bookends Pro is a **bibliography management system** available from Westing Software for scientists and other professionals who need to track and cite reference information for research and publication purposes (*Reader Service No. 114*). Designed for the Macintosh, the US\$149 Bookends Pro program offers a means of collecting, managing and formatting references, citations and quotes, and automatically generating bibliographies and footnotes that conform to publication specifications. Users can import reference information from a variety of online database services, including Medline, Dialog, BRS and others, as well as from CD-ROMs and tab-delimited text files. Over 90 predefined formats for generating references and foot-

notes are included; users may also customize their own. Bookends Pro includes fields that can be customized for authors, title, journal, volume, pages, publisher, keywords, notes and abstracts. Search and find options for locating records include searching on any field or combination of fields, boolean 'and', 'or' and 'not' criteria, unique or relative reference numbers, partial or whole word searches.

■ DNA/protein

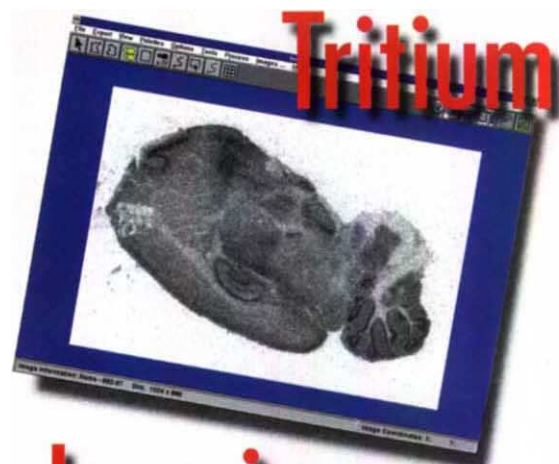
GeneWorks 2.3 is a Macintosh-based sequence analysis software package available from IntelliGenetics designed to **generate full-colour circular maps of multiple nucleic acid analyses** (*Reader Service No. 115*). With this latest upgrade, users can create maps of open reading frames (ORFs), sequence features and restriction enzyme digests, and view these analyses separately or in combination. Composite circular maps



Generate circular maps with GeneWorks.

make it easier to identify regions for designing vectors, subclones, hybridization probes and for mutagenesis studies. GeneWorks offers a range of analyses for DNA and protein sequences, including multiple sequence alignment, PCR primer design, dot-matrix sequence-similarity plots, sequencing gel assembly, sequence and translated sequence database searching, PROSITE motif and pattern searching, and protein secondary structure prediction.

EuGene is the new computer program from Daniben Systems for **designing oligonucleotide primers** for DNA amplification, sequencing and hybridization (*Reader Service No. 116*). This software for IBM PCs or compatible computers determines the best oligos for a particular task by considering a range of oligo lengths and melting temperatures, as well as maximum AT run lengths and GC content, hairpin stem sizes, palindromes, primer-dimer formation and 3'-stability. In addition, the user can rate the importance of up to eight oligo characteristics to make the program design oligos tailored to a specific application. The software will also check for uniqueness in the target DNA sequence, search for homology



Imaging

PhosphorImager™



Image your tritium-labeled samples with the Molecular Dynamics PhosphorImager™ system. New tritium-sensitive screens manufactured by Kodak are now available.



HIGH RESOLUTION

With 50µ pixel resolution get all the detail you need for whole body autoradiography. Use the PhosphorImager for all your radioactive samples — ³H, ³²P, ³⁵S, ¹⁴C, ¹²⁵I, and more.

ImageQuaNT™ analysis software and your choice of Macintosh or PC computer make the PhosphorImager system the most widely used instrument for autoradiography. Call today 800 333 5703.



Innovative Instrument Systems for the Life Sciences
928 East Arques Avenue, Sunnyvale, California 94086-4520 USA

Worldwide Contact Info: In France, (33) 1-6086-6513.

In Germany, (49) 2151-83870. In the U.K. and other European countries, (44) 732-762565. In Australia, (03) 810-9572. In the U.S. 408-773-1222, fax: 408-773-1493

©1994 Molecular Dynamics. All rights reserved. PhosphorImager is the trademark of Molecular Dynamics. Kodak is a trademark of Eastman Kodak Company. All other brands, products or services mentioned are the trademarks, servicemarks, registered trademarks or registered servicemarks of their respective owners. Product specifications are subject to change without notice.

in files of repeat sequences, evaluate a predesigned oligo, choose a companion oligo for PCR that will work with an existing oligo, and allow the addition of restriction enzyme sites to the oligos it designs.

Integrated Separation Systems has introduced new RFLPscan **software for DNA analysis** (Reader Service No. 117). This Windows-based software module can be used for the analysis of DNA restriction fragment patterns and other DNA gel images. The software can be used to process images obtained from a variety of sources, including laser scanners, flatbed scanners, phosphorimagers, CCD cameras and hand-held scanners.

For researchers **performing comparative sequencing**, Perkin-Elmer has introduced Sequence Navigator software (Reader Service No. 118). This new Macintosh-based



Navigator software for DNA/protein sequence comparisons.

program aligns sequences to a standard or a population and identifies sequence variants. It is designed to operate as part of a complete sequencing system, which includes the company's Prism sequencing reagents, Catalyst molecular biology labstation and Model 373 DNA sequencer. With this program, users can perform basic sequence editing, use pairwise alignments to assemble contigs, select multiple alignment to identify sequence variants, and display the electropherograms collected by the automated sequencer.

■ Hardware

Adept Scientific is introducing Strawberry Tree's alternative to the conventional plug-in board and external terminal panel for PC-based data acquisition. DataShuttle is a small **plug-in peripheral device for data acquisition**



Plug-in peripheral for data acquisition.

sition that accepts inputs directly from sensors and has an output that plugs directly into the parallel port of a PC (Reader Service No. 119). It eliminates the need for internal data acquisition boards, and should be of interest to those who want to use notebook PCs, or who need to collect data on desktop PCs where it is inconvenient to fit internal cards. Quicklog PC software comes bundled with the system. Prices start at UK£895.

Offering a peak performance of 1,000 million floating point operations per second and at a cost of under UK£50,000, the Paramid GF1 **desktop supercomputer** from Transtech Parallel Systems is designed to provide an affordable parallel supercomputing resource for researchers who find their existing UNIX workstations too slow for computing numerically-intensive applications (Reader Service No. 120). Such applications may include computational physics, finite element analysis, image processing and recognition, fluid dynamics modelling, seismic and signal processing, artificial intelligence and neural network applications. The Paramid GF1 uses a distributed memory architecture: ten processing nodes are used, each based around an Intel i860XP-100 MFLOP vector processor. Each multiprocessor has its own memory (16 MB per node) and communications processor with four communications links to other nodes, so performance is said not degraded by the need to access shared memory. The GF1 supercomputer connects to a UNIX workstation via a standard SCSI II interface.

IBM offers the scalable POWERparallel Systems SP2, **RISC-based parallel computer systems** said to combine the numeric-intensive processing capabilities of scientific and technical computers with the storage and analysis strengths of commercial systems (Reader Service No. 121). Using processor nodes based on IBM's POWER2 RISC technology, the new systems are designed to offer up to twice the node performance of previous systems, eight times greater memory and four times greater bandwidth. The SP2 can grow to 128 nodes, with a stated peak performance of 34 gigaFLOPS (equal to billions of calculations per second), 256 gigabytes (GB) of internal memory and 1,024 GB of internal disk storage. Suggested applications — seismic analysis, molecular modelling, computational fluid dynamics and plasma physics.

Last month, Silicon Graphics introduced POWER Onyx, a new **graphics supercomputer** based on the 64-bit MIPS R8000 RISC processor (Reader Service No. 122). Targeted disciplines include computational chemistry, oil and gas research, molecular modelling, global weather modelling, structural dynamics, image processing and animation. The POWER Onyx system, which is said to deliver up to 3.6 gigaFLOPS (GF) of peak performance, is offered with the com-

pany's RealityEngine² graphics subsystems, providing users with an interactive and high-end three-dimensional design and real-time capabilities. These subsystems contain 1.2 GF of floating-point processing power dedicated to performing geometric transformations. Available this month, the POWER Onyx system is supplied in configurations supporting two to 12 processors. Prices range from US\$88,000 to \$703,800.

Sun Microsystems introduces 24-bit **graphics workstations** (Reader Service No. 123). Four new 24-bit colour graphics SPARCstation 5S24 configurations, including a video conference-ready multimedia version, were added to the recently introduced SPARCstation 5 system line-up. Prices start at US\$6,595 for a 70-MHz system with a 17-inch colour monitor, 16 MB RAM and a 535 MB hard disk to a \$11,395 85-MHz system with a 20-inch colour monitor, 32 MB RAM and a 1 gigabyte hard disk. This line supports various 24-bit graphics applications, including Adobe's Photoshop. □

These notes are compiled by Diane Gershon from information provided by the manufacturer. For more details, fill in the reader service card bound inside the journal. Prices quoted are sometimes nominal and apply only within the country indicated.

* Molecular Mechanics * Quantum Mechanics
* Structure Predictions * IR / UV Visible Spectra
Powerful Molecular Modelling Software
easy to use & understand

HyperChem

for Windows Release 4

LIGHTWAVE Scientific
Tel: 0823 681006 (UK Only)
HYPERCUBE Inc
Tel: +1 519 725 4040

Reader Service No. 3

LIPOSOMES

REPRODUCIBLE — HOMOGENEOUS —
EXTREMELY RAPID



Our new LIPOSOMAT guarantees rapid (5ml within 60 minutes) preparation of uniformly sized unilamellar liposomes. Liposome size is selectable between 25 and ca. 600 nm diameter. Sample volume is normally 5-10ml or up to 200ml if used in combination with an additional instrument. Lipid concentration can be up to 300 mg/ml.

DIANORM

P.O. Box 650126, D-81215 Munich

Reader Service No. 4

Quality standards and the job market

J. Coombs

Adoption of quality standards, increased legislation and the need to prove due diligence are generating a growing market for personnel with skills in quality control, quality assurance and regulatory matters. This assessment comes from a company offering recruitment services in the field.

NOT long ago, the main worry for many small UK businesses apart from turnover and profit, was VAT (value-added tax) and PAYE (Pay As You Earn — the system under which employers deduct tax from salaries). Now, in addition, they are increasingly faced with a growing number of performance standards such as BS5750, BS7750 and ISO9000, which may be imposed by their clients. For larger companies these and other abbreviations* such as HACCP, TQM, GMP, GLP and NAMAS are commonplace.

The overall objective is to improve quality and to show how such quality is maintained. The concept of quality extends from the product itself through the working environment to the local or even the global environment. Within the European Union (EU), the pressure for change comes from existing and proposed directives, on packaging, on food additives, on pesticides and many more. They may lead to the introduction of QUID, possibly duplicating the situation in the United States where food labels grow in size and detail, legislated down to the point size used to define the nutritional value of a single serving contributing to a defined 2,000-calorie-per-day diet. At the other end of the scale, the consumer has become more demanding, requiring assurance that food is pure, fresh and free from additives and preservatives.

The pharmaceutical and medical sectors have a long history of regulation testing and trials. However, within the food industry, significant changes are being imposed, not only by legislation but also on smaller companies as a result of changes in the structure of the industry. Supermarkets are now competing with each other on service and quality, putting pressure on suppliers who pass it on down through the manufacturing and production chain. At the same time, the farmer is not helped by the adverse publicity arising from the development of bovine spongiform encephalopathy (BSE) in cattle, or the continued increase in cases of food-borne salmonella. Elsewhere, consumers respond adversely to revelations that 'pure orange juice' may be adulterated,

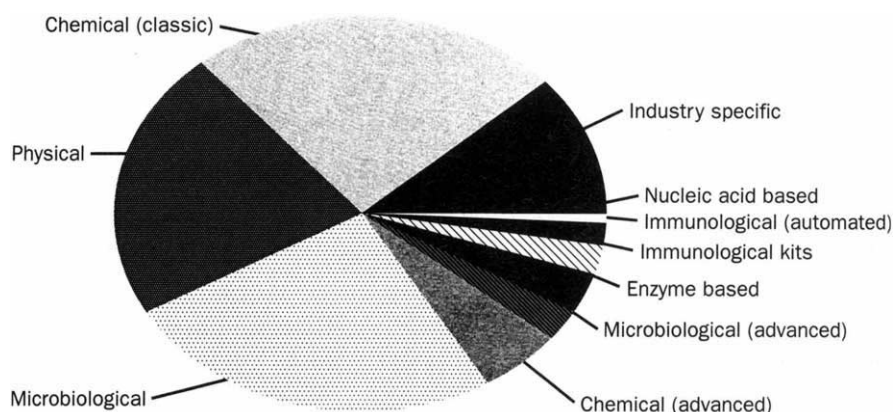


FIG. 1. Analytical techniques recorded during a regional survey of the food industry in the UK (frequency of a given response expressed as a percent of all responses).

that the term 'sugar-free' may be used to hide the presence of sugars other than sucrose and that the protein content of meat products may be of variable origin, extending to soya.

In August, the UK government proposed extensive guidelines on nutrition and health. Although these recommendations cannot be enforced, increasing legislation in other areas aims to prevent both disease and fraud, thus providing a healthy diet and good value for money to informed consumers.

It is up to companies to comply with this legislation and manufacturers to ensure that their products reach the quality demanded by retailers and ultimate consumers. In order to meet their obligations with due diligence, companies have to establish the ways and means. This requires procedures, sampling protocols and appropriate analytical methods.

During a regional survey of the agro-food industry in the United Kingdom, CPL Scientific Ltd found that a number of small companies have yet to address all these problems of quality. For example, visits to a number of smaller companies revealed few laminar flow cabinets, although most were carrying out microbiological plate counts. In many cases, it was clear that better quality control would be needed if the companies were to survive, yet the exact requirements may be unclear and costs high.

Companies must keep abreast of current and developing legislation, comply with competition laws, guard against adulteration (accidental or deliberate), screen raw materials against microbial contamination or aflatoxins, use appropriate

packaging materials and ensure that new product formulations do not generate hitherto unexpected results.

Companies, and specifically their directors, bear responsibility for compliance, maintenance of quality, enforcement of standards, avoidance of pollution of the environment, safe transport and disposal of wastes, provision of suitable storage and transport facilities and so on. In many cases, the proof of diligence lies in acquiring accurate analytical results from raw materials, manufactured products or the environment in which they are generated. The techniques used are biological (microbial plating, ATP-luminescence, enzyme assays, immuno assays), chemical (gas chromatography, high performance liquid chromatography, spectroscopy) or physical.

These aspects of analysis were also addressed in CPL's survey of the food industry, covering brewing, meat and poultry, fish, dairy, baking, flour and animal feed industries. Reasons for carrying out analysis were variously given as (in order of frequency of response) quality control, research, quality assurance, hygiene, due diligence, customer requirement, standards (BS, ISO, NAMAS), product development and technical audit. These tests covered raw materials, processes, products, equipment and staff swabs for hygiene, validation, inter-laboratory standardization, registration needs, waste and effluents as well as shelf-life studies. The relative importance of the various techniques are illustrated in Fig. 1.

Techniques are still dominated by the traditional microbiological, chemical and physical methodologies. As yet, in spite of

*Abbreviations: BS (British Standard), GLP (good laboratory practice), GMP (good manufacturing practice), HACCP (hazard analysis critical control point), HNC (Higher National Certificate), HND (Higher National Diploma), NAMAS (National Measurement Accreditation Service), QUID (quantitative ingredient declaration), TQM (total quality management).

their potential, advanced bioanalytical techniques (with a few exceptions such as the ATP-luminescence assay) based on immunology and gene probes are not in widespread use. In part this reflects cost relative to the narrow margins seen in the food industry. But another consideration is whether in fact such novel methods will be accepted by expert witnesses as satisfactory discharge of due diligence if it comes to a court of law. This is especially true where it is the level (or number of bacteria) that matters, not proving complete absence. This level may be prescribed in terms of colony forming unit, in a manner difficult to equate with the depth of colour response seen in an alternative assay.

Irrespective of the approved or chosen analytical method, to discharge this responsibility a company must hire someone with the required skills based on appropriate training and practical experience. This means it must select well qualified people and expect to pay a commensurate salary. In order to assess the availability of and requirement for such staff, as well as to identify suitable remuneration packages, CPL carried out an analysis of the characteristics of people applying to register on the national database run by its recruitment division. These candidates are mainly people in work looking for job advancement. CPL has also analysed the requests for staff from companies seeking to fill vacancies in the areas of quality control and legislative compliance. As a result, CPL is able to provide a profile of supply and demand. This covers qualifications and salaries relating to various levels of responsibility as well as some indication of where skill shortages may be occurring. These figures, covering applicants over a three-month period in the summer of 1993 and job requests for the first six months of this year, in general represent individuals in the 25–35 age group looking for their second or third job move. New graduates tend to be excluded as most employers require some previous industrial experience in these rather crucial areas.

In a recent survey of CPL's clients, more than 70 per cent listed relevant technical ability as the most important

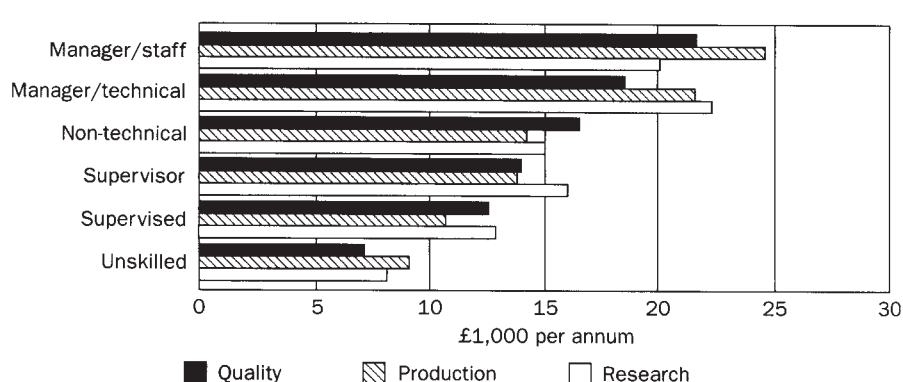


FIG. 3 Average salaries of workers seeking new jobs in life science industries.

criterion in selection. As a result, almost everyone looking for a job move into this area must have previous experience and appropriate scientific qualifications. Obviously, their qualifications differ with the level of responsibility. However, even the lowest grades are likely to have a BTEC. Those carrying out supervised work will require at least HNC (Fig. 2) as will those involved in decision-making in respect of quality control/quality assurance (QC/QA) matters, but not involved directly in laboratory work (indicated in the figure as non-technical decision-makers). Managers and laboratory supervisors are more likely to have a degree, but PhDs are apparently less common. In fact, this pattern may reflect the greater extent of applied training in what were the polytechnics, rather than the more academic environment of the traditional universities.

A prime consideration for many in considering a job or a move from one company to another, is what does it pay? The range of salaries of those seeking to move is shown in Fig. 3. For the sake of comparison, the levels for those involved in quality control and assurance are compared with the average salaries for those with similar technical qualifications holding jobs in product manufacture, product development or research are also shown.

These results are, of course, quite specific, reflecting the characteristics of those who choose to move, or, during a time of

quite dramatic restructuring in some industries, need to find a new job and see CPL Scientific recruitment services as a means of realizing this objective. However, with an overall sample size in excess of 1,500, these figures may be assumed to have some validity. They have been matched against the range of salaries on offer from organizations coming to CPL with vacancies. These have averaged around £20,000 for managers, close to the salary levels illustrated here. But offers for analytical staff, whether supervised or supervisory, are currently higher than those illustrated (£11,000–£16,000 and £15,000–£18,000 respectively). This may reflect a situation of increasing demand. Recently, more than a quarter of job requests received by CPL have been for such positions.

If this analysis is extended to cover those with the knowledge and ability to make informed decisions about regulatory affairs, then the apparent imbalance between supply and demand appears more extreme. Only a handful of experienced people with such background have approached CPL looking for a job change. In contrast, more than 26 per cent of the QC/QA job vacancies brought to CPL's attention request such skills, with salaries up to £30,000.

Overall it is clear that increased legislation and concern about quality are creating new opportunities for life scientists and chemists. However, to take advantage of these requires both relevant applied training and subsequent experience. In some areas, especially those related to legislative requirements, it would appear that shortages may already exist.

Quality-related analysis may be routine to a large extent, but it requires meticulous care and attention to detail as mistakes can jeopardize a company's reputation and possibly threaten human lives.

J. Coombs is technical director of CPL Scientific Ltd (Newbury), a private company offering technical consultancy, information and recruitment services in the life sciences.

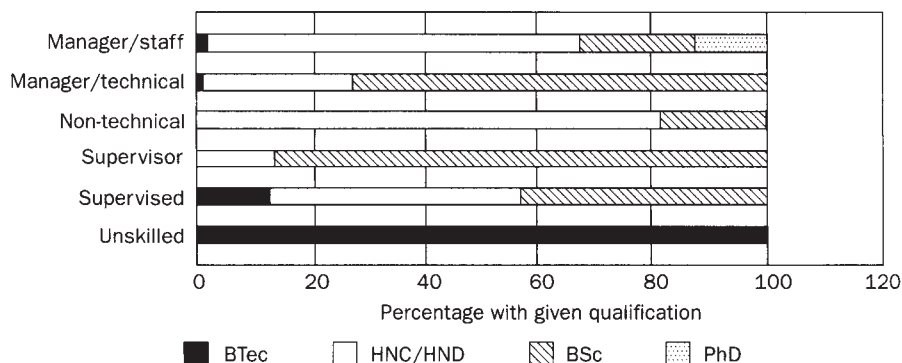


FIG. 2 Highest qualification of workers seeking new jobs in quality control and related areas.